The Moral Brain Hypothesis:
Neuro-psychological foundations of moral values and norms

Dissertation
submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Neuroökonomie, Dr. sc.
(corresponds to Doctor of Neuroeconomics, PhD)

Presented by

Giuseppe Ugazio
From Italy

Approved in September 2018 at the request of
Prof. Christian C. Ruff, PhD
Prof. Todd Hare, PhD

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 19.09.2018

Chairman of the Doctoral Board: Prof. Dr. Todd Hare

Abstract

Moral codes of conducts are a hallmark of human civilization: moral norms have been developed by every known human society to regulate social interactions. The emergence and maintenance of these norms has been proposed to reflect the evolution in the human brain of neurocognitive processes beneficial for cohabiting in large social groups, a perspective fully captured by the Social Brain Hypothesis. To date, however, there is still little knowledge of what these processes are and how they interact with other-purposes processes, such as those involved in (non-moral) value-based decisions.

In the present thesis I propose to close this gap by asking: are there neurocognitive processes that specifically represent moral values (rather than universally moral and material values)? How and where are moral and material value representations integrated to inform decisions? To address these questions, I conducted three studies where I combined behavioral measures of moral choice and preferences with neuroimaging (i.e., functional magnetic resonance imaging, fMRI) and brain-stimulation techniques (i.e., transcranial direct current stimulation, tDCS) to investigate the neural mechanisms involved in a) estimating the right course of action in morally ambiguous situations (i.e., where morally right and wrong behaviors depend on the agent's moral preferences), and b) implementing behavioral control functions that instantiate the behaviors prescribed by a moral norm shared by a group.

In a first fMRI study, I developed a novel experimental paradigm where I concurrently elicited subjective values in different choice contexts, a moral context and a financial context, in order to disclose a) if moral subjective values could be reliably estimated and used to explain individual moral preferences, b) if and where these subjective values are represented in the brain, and c) if subjective moral values are estimated by domain-general valuation processes or if these rely (to some extent) on moral-specific valuation processes. My results revealed that moral subjective values could be solidly estimated adopting traditional

computational models of choice processes. However, I found no evidence for a common neuro-computational process that estimated both moral and financial subjective values. Instead, the evidence suggested that computations of moral subjective values rely on domain-specific neural functions performed predominantly in the right temporo-parietal junction (rTPJ).

The second study used tDCS to alter the excitability of the right dorso-lateral prefrontal cortex (dlPFC) during acquisition of several behavioral measures, which allowed me to directly test a causal role for this brain area in arbitrating between self-interested motives and moral motives – in this specific case honesty. The data obtained in this context revealed that increasing excitability of the right dlPFC lead to a significant increase of honest behavior. No effect of tDCS was observed in a variety of other control tasks, suggesting a specific involvement of this brain region in regulating internal conflicts between selfish and moral motives.

Finally, the third study of the present thesis combined online tDCS and fMRI to examine at the neural-network level the causal brain activity responsible for implementing fairness, the moral norm that regulates how resources should be shared among members of a society. Intriguingly this approach allows me investigate how one brain area – the right DLPFC – may change how it influences whole networks in a context-dependent manner. Replicating a previous own study, behaviorally I showed that increasing/decreasing the excitability of lateral prefrontal cortex (LPFC) lead to an increase/decrease of compliance with the fairness norm in the presence of a credible punishment threat and to decreased/increased compliance in the absence of such punishment threat. These behavioral effects were mirrored by changes at the neural level. My fMRI results revealed that anodal stimulation of the right LPFC resulted in increased amygdala activity and increased connectivity between the stimulated area and the orbitofrontal cortex (OFC). In contrast, cathodal tDCS to the right LPFC induced

changes in functional activity in the anterior cingulate cortex (ACC), left LPFC, and bilateral inferior parietal lobules (IPL). Jointly, these findings support the hypothesis that the LPFC drives fairness-norm compliance by regulating the salience of inputs from neural areas processing social emotions – responsive to the threat of sanctions for norm violations – and cognition – strategically arbitrating when it is potentially safe to violate the fairness norm.

In my thesis I used morality, one of the most important tools regulating social interactions, to add evidence in support of the Social Brain Hypothesis. This hypothesis proposes that the cortical enlargement of the human brain mainly reflected the evolution of neuro-cognitive processes beneficial for co-habiting in large social groups. Taken together, my findings revealed some of these social-specific processes: study one identified neural activity specifically involved in the estimation of subjective moral values, compared to subjective financial material values; study two revealed evidence demonstrating a causal role for the right LPFC in arbitrating conflicts between competing moral and material values to determine moral behavior; finally, study three revealed that the LPFC controls moral behavior by selectively modulating the behavioral relevance of social emotions and cognition depending on the context in which moral decisions are taken.

List of Manuscripts

The dissertation is based on the following research articles:

Study 1:

Ugazio, G.*, Grüschow, M.*, Tobler, P. N., Lamm, C., Polania, R., and Ruff, C.C. The Neural Computations of Subjective Moral Value. (In Preparation)

Study 2:

Cohn, A., Marechal, M., Ugazio, G., and Ruff, C. C. (2017) Increasing Honesty in Humans with Electrical Brain Stimulation. *PNAS.*

Study 3:

Moisa, M.*, Ugazio, G.*, Hill, C., and Ruff, C. C. tDCS Induced Changes in Neural Networks Modifies Norm Compliance. (In Preparation)

Table of Contents

1. Introduction

A hallmark of humanity is the extent to which we rely on elaborated social codes of conduct - such as moral and social norms (Bicchieri 2005; Elster 1989; Axelrod 1984) - to regulate behavior in a social environment and peacefully resolve frictions generated by having to compete for resources with other members of a society (Tomasello 2011; Dunbar & Shultz 2007; Barton & Dunbar 1997). All known human societies, from simple hunter-gatherer groups to complex modern nations, use social norms to regulate intergroup interactions. Importantly, the need to develop emotional and cognitive skills that allow control of spontaneous behavior as well as the maintenance and development of such regulatory systems are thought to be a major force driving the evolution of the human brain, an assumption often referred to as the "Social Brain Hypothesis" (Dunbar 1998; Dávid-Barrett & Dunbar 2013).

Within the scope of this hypothesis, previous research has highlighted the relevance of three components for social behavior: social emotions (Reeck et al. 2016), i.e., emotions that are experienced vicariously (such as empathy or compassion) or emotions that result from imposing externalities on others (such as guilt or shame); social cognition (Apperly 2008; Samson & Apperly 2010; Hamilton 2005; Gordon 1996), mainly the ability to take the perspective of others or to attribute intentionality to the consequences of others' actions -- often jointly referred to as theory of mind, ToM (Saxe 2009); and social behavior control mechanisms that are at least partially independent of basic emotions and cognitions, for instance the control of selfish impulses or compliance with social norms (Montague & Lohrenz 2007).

These three aspects (emotions, cognition and behavior control) of social behavior have been proposed to play a fundamental role also for a sub-domain of social behavior, namely moral decision-making, the topic at the core of the present thesis. The next sections

review the literature describing some of the important facets of the moral decision making processes, illustrating what is known to date and, more importantly, what issues seem to still be unresolved and will be addressed by the three studies (see appendices to Studies 1-3) constituting the experimental part of this thesis.

1.1. Moral Decision Making

The origins of modern moral psychology can be traced back to Hume's Treatise on Human Nature (Hume 2000), where he proposed that moral decisions closely follow the emotional reactions that humans have in order to evaluate the moral appropriateness of an event. In the past century psychologists (Piaget 1932; Kohlberg 1971) proposed a more sophisticated developmental theory describing the different components of moral decision making and their stages of development in humans, from childhood to adulthood. One of these theories (Kohlberg 1976) proposed six moral development stages divided in three levels: pre-conventional, conventional and post-conventional. In the pre-conventional stages, moral reasoning is merely determined by self-interest (e.g., how can I avoid punishment), while in the conventional stages moral reasoning is motivated by interpersonal interests (e.g., how can I contribute to maintain social order). Finally in the post-conventional stage, moral reasoning takes a more metaphysical stance being concerned with universal ethical principles. Importantly, these scholars gave origin to two school of thoughts that disagreed on whether morality was primarily of emotional nature (Hume 2000) or of reason (Piaget 1932; Kohlberg 1971).

More recently morality has become the object of study of researchers from several disciplines including psychology, neuroscience, biology, psychiatry, and economics (Haidt 2001; Haidt 2012; Prinz 2006; Greene & Haidt 2002; Greene et al. 2004; Greene et al. 2001; Mikhail 2009; Shenhav & Greene 2010; Hauser 2007; Cima et al. 2010; Huebner et al. 2008; Harlé & Sanfey 2010; Sanfey 2007). The cumulative body of evidence from these disciplines

has demonstrated that morality is constituted by both affective and cognitive processes (Patil & Silani 2014; Majdandžić et al. 2012; Ugazio et al. 2012; Moll et al. 2008; Moll & de Oliveira-Souza 2007; Moll et al. 2005; Moll et al. 2002; Schnall, et al. 2008; Valdesolo & Desteno 2006; Wheatley & Haidt 2005; Greene et al. 2001; Shenhav & Greene 2010; Kliemann et al. 2008; Young & Saxe 2008; Cushman et al. 2006; Greene 2015; Young et al. 2007; Greene 2007; Greene et al. 2004; Cushman et al. 2010; Greene & Haidt 2002; Cushman & Young 2011).

Based on this body of evidence, several diverging models of moral decision making have been proposed. The main source of disagreement between these models is the relevance attributed to emotional and cognitive processes: some predict a more predominant role of emotions (Schnall, et al. 2008; Haidt 2001; Prinz 2006; Wheatley & Haidt 2005) while others propose that cognition and reasoning have a pivotal role in steering moral decisions (Mikhail 2009; Hauser 2007; Kohlberg 1971; Kohlberg 1976; Piaget 1932). To date, the most favored theory in this field is the dual-process theory proposed by Greene and colleagues (Greene 2015; Greene et al. 2004; Greene et al. 2001), which suggests that cognitive and emotional mechanisms compete in determining our moral judgment.

Some recent theories and studies in the moral domain have started stressing the importance of approaching moral decision-making from a value-based perspective (Crockett 2016; Crockett 2013; Ayars 2016; Cushman 2013; Shenhav & Greene 2010; Hutcherson et al. 2015), an approach already successfully used to explain choices and understand preferences in other decision-making fields such as economics (Clithero & Rangel 2014; Schultz 2006). To date, the few studies investigating moral decisions through the lens of value-computations proposed that these are related to both emotional and cognitive processes (Hutcherson et al. 2015; Shenhav & Greene 2010). However, how exactly these processes participate in the computation of moral values is still unclear. Similarly, there is still no solid

evidence that allows to determine whether moral decisions are represented in the brain by moral-specific neural processes or by domain-general mechanisms as one study proposes (Shenhav & Greene 2010). This proposal contrast with the possibility that the human brain may have evolved processes specifically dedicated to moral decisions, a view consistent with the Social Brain Hypothesis(Dunbar 1998; Dávid-Barrett & Dunbar 2013). More in detail, the Social Brain Hypothesis posits that humans have evolved neural processes that are particularly sensitive to social information: These processes thought to be selectively engaged in supporting and driving social behaviors and decisions, and are either not involved or involved to a lesser extent in implementing behaviors and decisions that lack this social component. For example (Spitzer et al. 2007; Ruff et al. 2013) demonstrated that the right DLPFC is selectively responsive to and causally engaged by compliance with social fairness norms. Crucially, in a decision situation closely resembling fairness norm compliance but lacking the social component, these region were significantly less engaged and was not found to causally determine non-social behavior. In this vein, study 1 compares the neural representations of social and non-social value computations demonstrating that there are also neural processes specifically responsible for performing moral value computations.

In the following chapters I discuss the available evidence indicating the existence of moral specific emotional, cognitive, and social behavioral control mechanisms.

1.2. Emotions and Moral Decisions

The idea that emotions have an important role in moral judgments was championed in Hume's works (Hume 2000). Building on these thoughts, psychologists and philosophers have more recently developed a series of studies testing the role of emotions in guiding moral decisions (Haidt 2012; Ugazio et al. 2012; Crockett et al. 2010; Harlé & Sanfey 2010; Schnall, et al. 2008; Valdesolo & Desteno 2006; Wheatley & Haidt 2005; Prinz 2006).

The first studies in this domain tested the hypothesis that emotions, by means of "gut feelings", inform a person about the moral correctness/wrongness of an action, by causing a given emotional reaction in the person experiencing a moral situation: a positive/pleasant emotion results from observing a morally praiseworthy event or a negative/unpleasant emotion emerges if such event is morally condemnable (Haidt 2012; Prinz 2006). This emotional reaction to a given moral event is then translated into a person's moral judgment: if this person is in a pleasant emotional state she will judge the event to be morally correct, whereas if she finds herself in a negative emotional state, she will judge the event to be morally inappropriate

To test this prediction, Haidt and colleagues (Schnall, et al. 2008; Schnall, et al. 2008; Wheatley & Haidt 2005; Haidt 2001; Haidt 2007; Haidt & Joseph 2004) developed a series of moral vignettes describing violations of moral norms that are suggested to be strongly connected to feelings of disgust, while at the same time controlling for some of the possible non-moral reasons that may be used to determine the appropriateness. For instance, one of the most representative scenarios describes two siblings who decide to have sexual intercourse but take measures to avoid pregnancy (Haidt 2001).

The results gathered in a series of studies seem to support the view that moral judgments in these situations are driven by emotions: the author reported (Haidt 2001) that participants expressing moral judgments in these vignettes did so by following their gut reactions rather than a reason-based moral principles. More precisely, participants reported that their moral judgments resulted from a disgust feeling evoked by the actions described in the vignettes. The link between disgust and moral judgment was further confirmed by several other studies (Schnall, et al. 2008; Wheatley & Haidt 2005) using different techniques to induce disgust including disgusting smells (Schnall, et al. 2008) and hypnosis (Wheatley & Haidt 2005). These studies reported that as a result of inducing disgust, participants' moral

judgments of these moral vignettes became harsher compared to the judgments of participants in a neutral emotional state.

Furthermore, this effect seems to be specific to disgust since no effect on moral judgments was found following the induction of sadness (Schnall, et al. 2008). Such specificity for disgust was further confirmed by a study (Schnall, et al. 2008) where purity, the opposite of disgust according to the authors, was primed in participants. In this case, purity-primed participants reported less harsh moral judgments on these vignettes compared to participants in a neutral emotional state.

While these studies used an experimental paradigm designed to specifically test the role of disgust for moral judgments, emotions have been shown to have an important role also in other less emotionally-salient moral situations, such as the trolley dilemma (Foot 1967) or the ultimatum game, an economic game used to test fairness-related preferences (Güth et al. 1982, for a review see (Güth 1995). The first of these relates to a scenario where one has to decide whether it is morally required to sacrifice the life of a person in order to save a large group of people, or vice-versa whether such action should be considered morally forbidden.

Using a version of this scenario known as the footbridge trolley dilemma (Thomson 2008; Thomson 1976), a novel study (Valdesolo & Desteno 2006) found that inducing happiness led participants to report more utilitarian moral judgments, i.e. considering it morally permissible to sacrifice one person to save a larger group. Further, another study (Ugazio et al. 2012) investigated the role of emotions for moral judgments taking into account not only the valence of the induced emotion – i.e. whether the induced emotion is a negative or a positive one – but also the motivational tendencies of the induced emotion – i.e. whether the induced emotion is an approach emotion or a withdrawal emotion (Berkowitz 2003). This study showed that inducing disgust or anger, two negative emotions of opposite motivational tendencies, resulted in opposing effects on participants' moral judgments: a

disgust induction resulted in a decrease of utilitarian decisions, while an anger induction resulted in an increase of utilitarian judgments pointing to a strong role for the motivational tendencies of emotions in shaping moral decisions.

The hypothesis that emotions affect certain types of moral decisions depending on their motivational tendencies was further corroborated by a study testing the role of emotions in fairness-based decisions (Harlé & Sanfey 2010). This study found that approach emotions, such as anger or amusement, lead participants to accept more often unfair offers compared to participants who were primed on withdrawal emotions such as disgust and serenity.

While the studies discussed so far focused on the relation between basic emotions and moral judgment, a more complex social emotion has also been shown to be an important factor influencing moral decision making: empathy (Yoder & Decety 2014; Patil & Silani 2014; Majdandžić et al. 2012; Ugazio et al. 2014; Crockett et al. 2010; Batson 2011; Gleichgerrcht & Young 2013). While it is still unclear how exactly empathy affects moral decisions, a proposed mechanism is that it decreases the willingness to endorse harmful actions. In line with this view, (Gleichgerrcht & Young 2013) found a negative correlation between the endorsement of utilitarian judgments in moral decisions similar to the trolley-dilemma and empathic concern: individuals with low empathic concern were more likely to indicate that the required moral course of action in these situations was harming one in order to save more.

A similar finding was reported by (Patil & Silani 2014) in a study on alexithymia patients. Furthermore, empathic concern was found to be an important modulator of the effect of serotonine-depletion in affecting moral judgments (Crockett et al. 2010). Consistent with the view that empathy affects moral decisions by increasing an aversion to harm, (Majdandžić et al. 2012) found that increasing empathic concern in participants led these to not endorse utilitarian moral judgments in trolley-type moral decisions.

The role of emotions for moral judgments was also studied at the neural level (Heekeren et al. 2005; Moll et al. 2002). In these two studies, the authors test neural activity elicited by stimuli presenting violations of moral principles as well as other emotionally charged stimuli of non-moral nature. The second study (Moll et al. 2002) revealed that attending to both moral and non-moral stimuli increased brain activity in the amygdala, the right thalamus, and the right insula/inferior frontal gyrus, brain areas associated with negative emotion processing. Importantly, contrasting the brain activity elicited by morally salient stimuli to the one elicited by stimuli without moral relevance the authors (Moll et al. 2002) found increased brain activity elicited by the morally charged pictures in the orbitofrontal cortex (OFC), the superior temporal sulcus (STS), and the medial frontal gyrus. These results suggest that moral aversive stimuli are processed by partially domain-specific neural functions.

Similar findings have been proposed in a study (Heekeren et al. 2005) measuring brain activity of participants while they read texts either describing moral norms violations (e.g., A gives B a bloody nose) or containing a violation of a grammatical norm (e.g., A dresses a very bloody wound). In this study the authors found that moral violations alone activated a network of brain regions which included the posterior STS, as well as the VMPFC, and the posterior cingulate cortex (PCC). Taken together, these studies revealed a dedicated functional network responsible for processing emotional reactions to aversive stimuli and stressed an important role of these neural processes in the integration of emotional reactions to moral evaluations.

Further neural evidence linking emotions to moral judgments can be found in Greene's works where he studied moral decision making in the context of the trolley-dilemma (Greene 2015; Greene et al. 2004; Greene et al. 2001). These studies analyzed the neural correlates of moral decisions by dividing moral scenarios into two classes: 1) personal

moral dilemmas, where sacrificing the life of a person is a direct consequence of the action judged and thus may elicit strong affective responses, and 2) impersonal moral dilemmas, where the loss of a life would result as a side effect of the action and thus presumably is associated with less affect (Greene et al. 2004; Greene et al. 2001).

The authors found that judgments in personal moral dilemmas that it is morally forbidden to sacrifice a life in order to save a larger amount of people relied on neural activity in brain areas associated with processing of emotions such as the Amygdala (Blair 2007) and in areas integrating emotional information into decision-making such as the LPFC or the vmPFC (Greene & Haidt 2002).

Based on the discussed studies, there is now ample evidence that establishes a strong role for emotions in moral decision making. More importantly, in line with the Social Brain Hypothesis, the discussed research gives strong indications that moral emotions are instantiated by domain-specific processes, represented at the neural level by brain activity in the OFC and STS among other areas (Moll et al. 2002; Heekeren et al. 2005). None of these studies, however, has investigated if and how these moral emotional processes participate in the more deliberative processes that underlie the intrinsic value of moral choice options, for instance the value of a human life. Consider one example of a common task used to measure moral judgments, the "footbridge dilemma" (Thomson 1976): A runaway trolley threatens to kill five people. The only way to save the five people is to push a stranger off a bridge, onto the tracks below. He will die if you do this, but his body will stop the trolley from reaching the others. Is it morally permissible to push this stranger off the bridge in front of the trolley? Existing studies suggest that, on average, approximately seven people out of ten judge killing the stranger morally inappropriate (Greene et al., 2001, 2004), even if this would lead to saving five people. This frequently observed refusal to endorse the harmful action has been suggested to result from a negative emotional reaction towards this action (Greene et al.,

2001, 2004, 2009). This in turn, suggests that emotions can modulate the value one assigns to the choice options considered. To date, however, despite over a decade of research investigating moral judgments in this dilemma context, very little is known about specific neuro-cognitive mechanisms that can explain individual differences in moral preferences, or their relation to emotional processes (for a review see Young & Koenigs, 2007). One study approached this issue analyzing the neural activity sensitive to the emotional attributes of a moral decision (Hutcherson et al. 2015). Here it was shown that, among other areas, the insula, and superior temporal gyrus correlated with emotional appraisals and that these representations were then integrated in an overall moral value judgment in the vmPFC. Importantly, this study also showed that emotional appraisals of moral decisions are only one of the components that concur in determining the value that a person assigns to moral choice options: in this study it was shown that other cognitive processes are involved in tracking other attributes of the moral decision, such as the objective utility of each of the choice options. For instance if we consider the moral dilemma above the number of lives that one could save/harm has been shown to be represented by different neural activity than the one underlying emotional appraisals, specifically in the right TPJ (Hutcherson et al. 2015).

Relying on the literature discussed in the present section, one of the aims of Study 1 of the present thesis is to generate new evidence that could clarify how moral emotions participate in moral value computations, and if individual differences in sensitivity to emotions could be used to explain differences in moral preferences among people.

1.3. Cognition and Moral Decisions

Several studies in moral psychology have recently stressed the importance of cognitive mechanisms in shaping our moral decision-making. Of particular relevance is the ability to infer and attribute intentionality to the actions of other people and of taking the perspective of others, a social cognitive mechanism often referred to as Theory of Mind (Saxe et al. 2004).

Testing the role of Theory of Mind in moral intuitions and decision making, Young and colleagues (Young et al. 2007) developed a new set of moral scenarios in which one had to determine whether a fictitious agent was morally culpable for his actions. These actions were described as resulting in consequences that were either negative (the agent harming the peer) or neutral (the peer was unharmed by the agent's actions). Furthermore, the intentions of the agents were also manipulated, being described as either negative (e.g., intending to do harm) or neutral.

Combining these two factors, moral judgments were analyzed in four different contexts: 1) where the intentions were neutral and the outcome neutral, i.e., no harm intended and no harm caused; 2) where the intentions were neutral but the outcome negative, yielding a situation where harm was caused unintentionally; 3) where the intentions were negative but the outcome neutral, yielding a situation where a person attempted to cause harm but failed; and 4) where the intentions were negative and the consequences negative, i.e., a situation where harm was intended and successfully caused.

The behavioral results reported in this study (Young et al. 2007) highlighted the importance of intentionality in moral judgments. As expected, participants considered actions leading to harmful consequences more morally wrong compared to actions that resulted in neutral outcomes. Importantly, intentions modulated these moral judgments of condemnation: unintended harmful actions were considered less morally wrong than actions that intentionally caused harm; similarly, actions attempting to commit harm were judged to be more morally wrong compared to actions that did not intend any harm.

Furthermore, measuring participants' brain activity while reporting their moral judgments in these moral scenarios, (Young et al. 2007) found that the modulatory effect of intentionality on moral judgments was predominantly represented in neural activity in the right TPJ, a brain area considered to play an important role in implementing Theory of Mind

(Saxe et al. 2004). In a subsequent study (Young et al. 2010), the authors used this moral decision-making task while disrupting right TPJ neural activity in participants using transcranial magnetic stimulation (TMS) to test if the right TPJ had a causal role in interpreting the intentions behind other people's actions and incorporating this understanding in moral evaluation of the observed actions.

This study found that when the TPJ was disrupted with repetitive TMS, participants based their moral judgments more on the consequences of the agent's actions rather than on his intentions, in particular for scenarios in situations with attempted harm. Taken together, these results revealed an important role of Theory of Mind in shaping human moral judgments, and provided evidence for a causal link between this ability and neural activity in the right TPJ.

Another cognitive process important for moral decisions is the one that enables us to compare magnitudes and estimating values. While in the previous section, I discussed the role of affective processes in personal trolley-dilemmas, the same studies highlighted an important role of cognitive mechanisms for driving moral judgments in impersonal moral dilemmas (i.e., where harm to a person would result as a side effect of the action). In this specific context, considering it morally permissible to sacrifice the life of a person to save more people was found to activate brain areas implicated with cognitive processes (e.g., working memory, (Cohen et al. 1997; Smith & Jonides 1998; Smith & Jonides 1997)), such as the parietal lobes (Greene et al. 2001; Greene & Haidt 2002; Greene et al. 2004), and value-based decision making (Shenhav & Greene 2010; Hare et al. 2008; Schultz 2006) such as the medial orbitofrontal cortex (mOFC), and the ventral striatum (VS).

Notably, one recent study (Shenhav & Greene 2010) investigated if computations of expected values in moral decision making elicited the same neural mechanisms typically found to represent economic expected values. The evidence reported in this study suggests

that moral decision making is largely supported by domain-general mechanisms involved in computing expected values: the authors found that expected-values computations in the context of moral decisions elicited neural activity, among other areas, in the mOFC and the vmPFC (Shenhav & Greene 2010).

Since these brain areas are the very same ones previously identified in a number of non-moral tasks as important for computing the expected values of given decision options (Platt & Huettel 2008; Knutson et al. 2005; Knutson & Peterson 2005), the authors propose that moral decisions rely on domain-general decision mechanisms rather than on moral-specific ones. In line with this view, recent works (Ayars 2016; Crockett 2013; Cushman 2013) have proposed to adopt social-learning mechanisms from the economic literature (Clithero & Rangel 2014) in order to explain how moral intuitions – i.e. the intuitions that constitute the foundations of our moral preferences – result from domain-general learning mechanisms and are shaped by social interactions.

Critically, however, the proposed domain-generality view relies only on reverse inference, since no existing study has directly compared the neurocognitive processes involved in the estimation of moral values to those involved in estimating other types of values. Study1 of the present thesis addresses exactly this issue, by directly comparing moral and financial value computations. Only through such a direct comparison it is possible to clarify whether moral value computations rely on domain-specific mechanisms, as one would predict from the perspective of the Social Brain Hypothesis, or on domain-general mechanisms.

The literature discussed in this and the previous sections has highlighted the existence of two different types of processes involved in driving moral choice, one relying more on affective components, the other on cognitive processes. Crucially, it is yet to be clarified to

which extent these affective and cognitive processes contribute to moral decisions through domain-general or moral-specific mechanisms.

As mentioned, the dominant view in the moral psychological literature (Greene 2015) proposes that these two mechanisms are at the basis for different moral judgments (e.g. utilitarian or deontological judgments). For instance, (Greene 2015) suggests that utilitarian moral judgments are mostly driven by cognitive processes, calculating and comparing the value of saving five people to the value of saving one person. However, deontological moral decisions are mainly influenced by a strong aversive emotional reaction towards the act of directly killing someone. Moral judgments therefore seem to be the outcome of a competition between the cognitive information that maximizing utility may be more desirable versus the emotional aversive reaction against violating a human right, such as the right to physical integrity, thus supporting the view to consider such action undesirable. Which of the two mechanisms prevails is proposed to be arbitrated by mechanisms of behavior control (Greene 2015; Greene et al. 2004; Greene & Paxton 2009) as discussed in the next section. It is important to note that recent studies approaching moral decision-making from a value-based perspective (Crockett 2016; Crockett 2013; Ayars 2016; Cushman 2013; Hutcherson et al. 2015) are casting doubts on this dual-process competitive inhibition account, generating evidence that instead supports the view that moral decision making resembles the architecture of other value-based decisions, e.g. economic choices. In this case, distinct regions represent emotional and cognitive decision variables independently, and are integrated into an overall value signal. In this case, therefore, conflict between decision variables does not necessarily imply dual systems competing for control, but may instead indicate competition in the processes computing the various decision variables.

Due to a substantial lack of evidence, it is unclear whether moral value-based decision making closely resembles other types of value based decisions. In this thesis, in particular in

Study 1, I propose to fill this gap with a study in which I can accurately characterize the moral value computation processes. With this approach I can therefore be in an ideal position to examine the neural correlates of the subjective values of moral choices, and then to contrast them with the neural correlates of subjective values of financial choices.

1.4. Behavioral Control and Moral Decisions

In order to perform several different types of tasks, the human brain relies on multiple neuro-psychological mechanisms. Each of these mechanisms is specialized to process different pieces of information. For instance, the fusiform face area (FFA) is specialized in processing inputs related to facial characteristics (Kanwisher & Yovel 2006; Kanwisher 2006), area V4 (possibly in combination with other parts of the visual cortex) of the visual system processes colors (Knoblauch 2002; Gazzaniga 2002), while Broca's area is specialized in elaborating language-related content (Fadiga et al. 2010; Fadiga et al. 2009).

In many cases, these different mechanisms operate in synchrony, allowing us to efficiently and accurately perform many types of tasks. In other situations, however, the inputs from these mechanisms yield contrasting responses, resulting in a conflict that hinders the accuracy and efficacy of our behavioral ability to solve a given task. One of the most famous examples of these situations is captured by the Stroop task (Botvinick et al. 2001; Stroop 1935; MacLeod 1991). This task entails reading out loud the color in which a word is printed. Importantly the semantic meaning of this word is itself a color. The performance in this task varies if the semantic meaning of a word and the color this word is printed in are matched (e.g. red printed in red ink) or are different (e.g. red printed in green ink).

In the first case, most people can perform the task accurately and rapidly, suggesting that the different neural mechanisms coding for the different properties of the stimuli are attuned in priming the same response (e.g. responding red in the given example). In the second case, however, not only the accuracy of the responses is lower than in the previous

case, but also giving a response takes much longer (van Maanen et al. 2009; Jensen & Rohwer 1966; MacLeod 1991).

The behavioral differences between the two tasks, known as Stroop Effect, have been shown to result from conflicts between systems processing information relevant to performing this task in the case where semantic meaning and printing color are incongruent. In this case, in fact, the task of naming the printed color receives competing information from the visual system carrying the color information (e.g. red) while it receives different information from the mechanisms processing the semantic meaning of the word (e.g. green).

In order to give a response, this conflict needs to be solved by behavioral-control mechanisms (Botvinick et al. 2001; Cohen et al. 1990). These mechanisms are responsible for mediating between the systems in conflict, downregulating one of the two systems so that a decision can be taken. The response will then rely on the information provided by one of the two (or more) systems triggered by the given task. In the example mentioned here, if the visual system prevails then the response will be 'red', while if the semantic information is carried on to the decision the response will be 'green'.

Behavioral control mechanisms have been proposed to be at play in resolving conflicts between several types of processes: from the basic cognitive conflicts described above to basic emotional conflicts (Dresler et al. 2009; Ochsner & Gross 2005; McKenna & Sharma 2004; McKenna & Sharma 1995; Williams et al. 1996), and more importantly in social-emotional (Silani et al. 2013) and social-cognition conflicts (Baumgartner et al. 2011; Knoch et al. 2008; Knoch et al. 2006; Knoch et al. 2009; Knoch et al. 2010; Knoch & Fehr 2007; Greene 2007). These studies have highlighted an important role of different neural structures and functions implementing behavioral control in these different situations, including the anterior cingulate cortex (ACC) and the LPFC for basic and social cognitive conflict resolution (Ruff et al., 2013; Knoch et al., 2009; Greene et al., 2004; Botvinick et al.,

2001), while the rTPJ has been linked to social emotional behavioral control (Silani et al. 2013).

Of particular relevance to this thesis is the role assigned to the right LPFC in behavioral control during various moral decision making tasks. As mentioned above, Greene and colleagues (Greene 2015; Greene 2007; Greene et al. 2004; Greene et al. 2001) have proposed that moral decisions in personal trolley dilemmas rely on conflicting decision systems. Evidence for the existence of multiple mechanisms competing for action control was revealed by increased reaction times required for a person to report their moral judgment in these situations - especially for judgments considering it morally permissible to sacrifice a person to save more - compared to the time it took to make moral judgments in impersonal moral dilemmas. This view was further supported by fMRI data revealing stronger activation in the ACC (Greene et al. 2004), one of the brain areas associated with decision-conflict resolution of the type described above (Botvinick et al. 2001; Shenhav et al. 2013).

While in the case of the trolley-type moral decisions the source of the conflict is suggested to be between affective, harm-aversion related processes and more cognitive processes coding the number of lives at stake (Greene, 2015), the need for behavioral control mechanisms has also been identified for choice-situations where participants have to choose between behaving selfishly (e.g. earning larger amounts of money) or behaving in adherence with the moral prescriptions applying to a given choice-situation (e.g. reciprocating trust or sharing monetary rewards with others fairly). Several studies (Baumgartner et al. 2011; Knoch et al. 2008; Knoch et al. 2006; Knoch et al. 2009; Knoch et al. 2010; Knoch & Fehr 2007; Greene 2007) have attributed a critical role for implementing behavioral control in these situations to the right LPFC.

For instance, (Spitzer et al. 2007) reported that this brain area responded strongly to the presence of sanction threats during exchanges in a modified dictator/ultimatum game. In

particular, its activity was increased when the participants had to choose how much money to transfer to opponents who could punish for norm transgressions, compared to opponents who could not react to the transfer. According to the authors, the neural activity in this brain area corresponds to increased behavioral control necessary to resist the temptation of behaving selfishly in order to avoid being punished by their counterparts.

In a subsequent study, (Ruff et al. 2013) used anodal/cathodal tDCS to increase/decrease the neural excitability of the right LPFC to test whether it indeed controls behavioral reactions to such punishment threats. To this end, they measured both voluntary norm compliance (transfers in a dictator game where the opponent cannot punish) and sanction-induced norm compliance (changes in transfers from the voluntary compliance when sanction threats are present, i.e., when the opponent can punish). The results revealed that both types of norm compliance could be modified with tDCS in opposite ways: sanction-induced norm compliance was increased by anodal tDCS and decreased by cathodal tDCS, whereas voluntary compliance was reduced by anodal tDCS and enhanced by cathodal tDCS. This pattern of results suggests that anodal/cathodal tDCS rendered participants more/less sensitive to the presence of sanction threats, causing larger/smaller adjustments of behavior to the external incentives.

Several other studies using brain stimulation to disrupt neural activity in this brain region (Baumgartner et al. 2011; Knoch et al. 2009; Knoch et al. 2006) reported results consistent with the view that the right DLPFC plays a critical role in social behavior control: following such disruption, participants were less able to build a positive reputation in a repeated interaction trust game (Knoch et al. 2009) or to reinforce fairness in ultimatum games (i.e. not punishing norm-violators, Baumgartner et al., 2011; Daria Knoch et al., 2006), when these choices required behavioral control of selfish impulses (e.g. monetary rewards).

Furthermore, some of the previously described studies suggest that the right LPFC is particularly involved in control of behavior in social contexts. The previously mentioned fMRI study by (Spitzer et al. 2007), for instance, identified socially-specific neural correlates of social norm compliance by comparing neural activity when participants interacted with another person versus when they performed the exact same task, but now against a computer algorithm. Importantly, the right LPFC that was later targeted with tDCS by (Ruff et al. 2013) was more strongly activated by the presence of sanction-threats when participants interacted with human opponents than with the computer algorithms.

Using a similar approach, (Ruff et al. 2013) also investigated the effects of LPFC tDCS in social vs. non-social contexts, finding that the strength of the tDCS effects were significantly more pronounced in the social condition. These findings indicate that the neural activity in the right LPFC is indeed causally necessary for behavioral control only during social interactions with other humans, rather than reflecting other aspects of the choice situation that may be similar for non-social choices (such as risk assessment, response selection, etc.). A similarly social-specific causal role of the right LPFC was also identified by (Knoch et al. 2006), where it was shown that TMS only increased acceptance of unfair offers from a human opponent but not from a computer.

In sum, this section reviewed studies that highlighted the importance of behavioral control mechanisms for moral decision-making. Specifically, the identified mechanisms regulate the relevance of emotional and cognitive elements for determining moral decisions. Critically, the studies discussed in the previous paragraph provide solid evidence for the existence of social behavioral control mechanisms in the right LPFC that determine moral decisions. Based on a growing body of evidence (Duncan 2001; Miller & Cohen 2001; Buckholtz et al. 2015) suggesting that the right LPFC has the function of integrating information from several interconnected neural networks, one possible mechanisms through

which this behavioral control function is exerted is through of the integration and modulation of information across different brain regions depending on different behavioral contexts. Study 3 tests exactly this hypothesis with the aim of clarifying how the neural function in the right LPFC relates to neural activity in other brain areas in order to regulate behavior.

Further, the evidence presented in this section adds to the data discussed in the section on emotions and morality (Section 1.2), in strengthening the view that the human brain recruits domain-specific mechanisms to implement moral decisions. The proposed hypothesis that the human brain evolved moral-specific decision mechanisms, which I support with fresh evidence in Study 1, opens a further intriguing question: how and where does the brain integrate inputs from moral-specific and domain-general processes? I will answer this question in study 2 assigning an important role for this function to the right DLPFC. Moreover, despite the evidence discussed in this section, it is still unclear how the brain activity in the right DLPFC changes its relation with other brain functions ultimately causing changes in moral behavior. This question is addressed in Study 3.

2. Summary of the Experimental Strategy

As illustrated by the extensive literature review above, research in moral psychology has yielded a sophisticated understanding of the latent principles that guide moral decisions (Haidt & Joseph 2004; Cushman et al. 2006; Gray & Wegner 2009; Mikhail 2009; Malle et al. 2014) as well as detailed maps of the neural substrates supporting these decisions (Greene et al. 2004; Young & Dungan 2012). To date, what is still lacking is a mechanistic understanding of how all the identified processes involved in moral decisions lead to differences in moral value computations, preferences, judgments, and behaviors. Achieving this mechanistic understanding could further contribute to disclose whether moral values are supported by domain-specific neurocognitive processes.

In the present thesis, I propose to generate such a mechanistic understanding using various psychological and economic measures in combination with correlative (i.e., fMRI) and causal (i.e., tDCS) neuroscientific methods. More precisely, I first use the literature on value-guided decision making to specify a computationally precise model to measure if a) moral preferences relate to neural activity in specific brain areas, and b) whether there is an overlap in the neural representations of moral preferences and other types of preferences, e.g. in the financial domain. Second, I use the literature on brain stimulation and social neuroscience to establish the causal role of the right LPFC in arbitrating between moral and material motives in determining behaviors. Third, relying on insights from all of the above literatures, I combine behavioral, neuroimaging, and brain-stimulation methods to investigate how (exogenously induced) changes in the right LPFC excitability affect moral behavior via the modulation of different processes in several brain regions.

In study 1, I built on the literature on moral dilemmas mentioned above (Shenhav & Greene 2010; Greene et al. 2004; Thomson 1976) as the inspiration for my task design. By definition, moral dilemmas do not have an objectively right or wrong answer, since any solution depends substantially on the meta-ethical principles a person believes in. These dilemmas thus provide an ideal theoretical framework to study how subjective moral preferences relate to subjective moral value computations, as well as the neural substrates underpinning such computations.

Studies 2 and 3 employed experimental strategies where the morally prescribed behavior is governed by a salient moral norm. Combining these experimental paradigms with tDCS, I could therefore study the causal involvement of brain activity in the right LPFC in determining behavior when the morally prescribed actions are challenged by competing selfish opportunities (e.g. increased monetary earnings). Study 2 relied on previous correlative evidence (Greene & Paxton 2009) implicating right-DLPFC neural activity in the

ability to resist the temptation of lying when doing so would result in increased monetary gains. Increasing or decreasing neural excitability in this brain area with anodal or cathodal tDCS, while participants performed a task measuring their compliance with honesty, it was possible to investigate if this brain area has a causal role in arbitrating between the conflicting (moral vs. material) motives.

Finally, study 3 combined behavioral measures of fairness with concurrent brain imaging and stimulation to determine the causal neural networks underlying fairness norm compliance. Building on a previous study where it was shown that modulating neural excitability of the right LPFC resulted in changes of fairness norm compliance (see above, (Ruff et al. 2013) I investigated if such behavioral changes were reflected by a tDCS-induced reorganization of the neural functions necessary to determine norm compliant behavior, as one would expect based on evidence suggesting that the behavioral regulatory function of this region operates via the modulation of interconnected brain networks activity (Duncan 2001; Miller & Cohen 2001; Buckholtz et al. 2015).

The results obtained in these three studies concur in answering some of the most pressing questions in moral psychology and have important implications more broadly for behavioral sciences such as social psychology, behavioral and neuro-economics, as well as philosophy. Specifically, I provide crucial evidence for the Social Brain Hypothesis, by demonstrating the existence of domain-specific moral value representations, by identifying a region in the right LPFC that is causally relevant for arbitrating between moral values and material rewards, and by generating a more fine-grained mechanistic explanation of how the right LPFC functionally interacts with other brain regions in order to drive moral behavior.

2.1. Study 1: The Neural Computations of Subjective Moral Value

Background

The neural mechanisms underlying moral decision-making have been extensively investigated with moral dilemmas (Greene 2015; Pascual et al. 2013) that require judgments about whether it is morally permissible to harm a smaller number of people in order to save a greater one (Thomson 1976). However, little is known about specific neuro-cognitive mechanisms that can explain why certain individuals judge harming the smaller group morally wrong while others judge it to be appropriate or even mandatory.

In the present project, I approach this issue from the perspective of the value that participants place on each of the lives under consideration. Computations of values for choice options play a crucial role in many other forms of decision-making (e.g., economic), and recent studies have proposed that comparable neural value representations may also underlie moral decisions (Shenhav & Greene 2010). Here I use fMRI and two closely matched decision tasks (a moral task and a financial task) to identify and directly compare the neural computations of subjective moral vs monetary values.

Methods

The moral task required participants to solve a dilemma similar to the classic footbridge dilemma: Should they sacrifice the life of one person in order to save several other people? In order to calculate subjective moral values, I manipulated two factors in the dilemma from trial to trial: The number of people (min =1, max = 10) that could be saved and the moral deservingness (criminal record; from none to mass murder) of the person that would need to be sacrificed. This latter factor allowed me to identify how much each participant "discounted" the moral value of a person's life based on his criminal record, in close resemblance to the well-known temporal discounting of monetary values. In the matched economic paradigm, I estimated participants' subjective monetary value in a standard temporal discounting task (McClure et al. 2007; Luhmann 2009; Green et al. 2004) entailing choices between 20 Swiss Francs immediately or variable larger amounts (min = 22, max =

120 Swiss Francs) at a later date (min= 1, max = 180 days later). The same standard hyperbolic discounting model was fit to each participant's behavior in both tasks, to derive subject-specific predictions for the moral and monetary values computed on every trial.

Formally, behavior in the moral task was modeled with the following hyperbolic function: SVm = 1 / (1 + Km * Deservingness), where SVm is the subjective moral value of saving the lives of the larger group by sacrificing the life of one person. The 1 in the numerator reflects the person one needs to sacrifice in order to save the larger group; Km corresponds to a subject-specific moral discounting constant, and Deservingness models moral deservingness (i.e. criminal prior records) of a given person. Similarly, behavior in the financial task was modeled with the following hyperbolic function: SVf = 20 / (1 + Kf * Delay), where SVf is the subjective financial value of the delayed option estimated as fraction of the immediate reward, 20 represents the immediately available option (i.e. 20 CHF), and Kf corresponds to a subject-specific financial discounting constant, and Delay models the time (in days) that people had to wait in order to receive the reward.

Results

Consistent with previous findings (Kable & Glimcher 2007), my participants' discounting curves were well modeled by a hyperbolic-discounting function (R2 = 0.98±0.015), revealing the expected variance of financial discounting factors Kf, and hence of the Subjective financial values (SVf) across participants (min Kf = 3.78*10-5; max Kf = 0.43404). Further, in line with previous evidence, SVf correlated with neural activity in the VS (small-volume correction, $P < 0.05$), as well as all in the vmPFC and PCC (threshold at $P < 0.001$).

Subjective moral values, SVm, were estimated using a structurally identical hyperbolic function. My participants' discounting curves were well modeled by the hyperbolic model implemented (R2 = 0.96±0.03) and the corrsponding moral discounting factors (Km) also varied across participants (min Km = 9.3*10-2; max Km = 7.083825). At

the neural level, I identified neural correlates of SVm in the bilateral Anterior Insula, the left inferior parietal lobule (IPL), and the anterior cingulate cortex (ACC), suggesting that stronger activity in these brain areas is associated with utilitarian moral judgments. In addition, this data also revealed that an increase in the likelihood of expressing non-consequential moral judgments (i.e., considering it morally forbidden to harm one person) correlated with BOLD activity in the rTPJ, the PCC and the right DLPFC, areas that I expected to be involved in moral valuation based on previous related studies (Kliemann et al. 2008; Young et al. 2007; Greene et al. 2001).

Comparing SVf and SVm, I could directly test if moral and financial value representations are implemented by similar psychological processes and represented in neural activity in overlapping brain structures. Importantly, while the two types of choices were comparable in terms of their computational requirements, they obviously differed qualitatively in terms of choice options and their consequences: On one hand, participants made decisions about whether or not to harm a human to save other lives, while on the other, they decided between different financial payoffs. It may therefore be expected that the two types of choices may differ in terms of response difficulty. However, response times (RTs) for the two types of decisions – a standard proxy to measure task difficulty – did not differ significantly (average RTs moral 1228.78ms +/- 257 (s.e.m.), financial 1226.25ms +/- 164 (s.e.m.), $t(21) = 0.04$, $p = 0.967$). Thus, the two types of decisions did not differ in the associated choice difficulty, making it possible to compare the underlying neural mechanisms without any possible confound due to differences in task difficulty.

At the behavioral level, I could not find a significant correlation ($r = 0.11$, $p = 0.59$) between the two types of discounting factors (Kf and Km). Taken with the due caution, an absence of correlation favors the view of distinct, rather than overlapping, processes implementing the two types of SV computations. More importantly, by directly comparing

the neural activity sub-serving the neural computations of SVm vs. SVf, I found that neural activity in the rTPJ and the PCC was more strongly involved in the computations of SVm than of SVf ($P < 0.001$). Critically, this provides for the first time evidence that moral values are represented by domain-specific neural activity in line with the predictions of the Social Brain Hypothesis, and contrasts the reverse-inference driven hypotheses proposed by previous studies.

Conclusion

This study provides critical evidence that advances our understanding of morality, by showing that subjective moral preferences (as captured by subjective values of moral options) are explicitly represented by neural activity in the brain. Crucially, the neuro-imaging results proposed in this study revealed that differences in moral preferences can be explained by differences in the neural functions elicited by the task: participants who had a more deontological moral preferences displayed stronger activity in brain areas associated with harm-aversion, emotional processing, and theory of mind. In contrast, participants with more utilitarian preferences displayed stronger activity in brain areas involved in conflict resolution such as the ACC or the left IPL. Furthermore, these moral values computations are clearly domain-specific, as they are spatially dissociated from computations of monetary values under comparable choice situations. This suggests that moral and purely monetary value computations may occur in parallel to drive human choices.

2.2. Study 2: Enhancing Honesty with Brain Stimulation

Introduction

Honesty plays a key role in virtually all human interactions and is of paramount importance for efficient social, economic and political institutions. Not much is known about the neural processes that enable humans to remain honest when tempted to lie. This is because lying or

cheating are inherently private acts coined by concealing the truth from others without their knowledge, in order to secretly increase one's own benefit. This contrasts starkly with previous neuroimaging studies on deception (Abe et al. 2014; Sip et al. 2008) that explicitly instructed participants to give wrong statements with the experimenter's knowledge and without any personal consequences emerging from such deceptive acts. Thus, it remains largely unknown which brain processes enable humans to stick to the truth when faced with the option to cheat in a truly concealed fashion. Here I show that honest behavior has a neurobiological basis in the right dorsolateral prefrontal cortex (DLPFC) and can be increased by transcranial direct current stimulation (tDCS) of this brain structure. It has to be noted, however, that since tDCS does not allow achieving highly focal stimulation effects it is impossible to make precise inferences on a clear localization of the neural processes governing honest behavior within this brain region.

Methods

I measured dishonesty with an innovative paradigm in which participants faced the temptation to cheat in order to increase their earnings without any danger of being detected and, eventually, sanctioned. Participants self-reported the outcome of a series of dice throws that had different financial consequences. Due to the truly concealed nature of the participants' behavior in my paradigm, dishonest behavior could be statistically detected at only at the aggregate group level, and not at the subject level.

One previous neuroimaging study (Greene & Paxton 2009) employing a related approach had identified a specific part of the right DLPFC that showed increased BOLD activity when participants successfully refrained from cheating. This finding provided correlational evidence that the right DLPFC is involved in decisions that require a moral trade-off between honesty and financial gain. Based on this finding, I exogenously increased neural excitability in this specific brain area with anodal transcranial direct current

stimulation (tDCS, 1.5 mA for 20 minutes, 49 participants) while participants performed a test battery containing the crucial task measuring dishonest behavior. This test battery contained also tests for various other choice processes such as risk and ambiguity preferences, impulsivity, or altruism that allowed me to examine if the behavioral control function I hypothesized for this region is specific to honesty, or is more broadly involved in regulating other types of choices putatively requiring some sort of behavioral control. I controlled for unspecific stimulation effects by also running two control groups in which DLPFC excitability was left unchanged (sham, 47 participants) or decreased (cathodal, 49 participants) by tDCS.

Results

Consistent with previous findings, participants in all three stimulation groups were dishonest, misreporting their results on an estimated 37% of trials. More importantly, however, I found that increasing right DLPFC excitability by means of anodal stimulation significantly reduced the percentage of successful dice rolls to a 15% rate of misreporting ($z=2.811$, $p= 0.005$, Mann-Whitney), corresponding to an approximately 60% lower cheating rate than in the sham and cathodal conditions. Note that these two latter conditions did not statistically differ ($z=-0.475$, $p= 0.6348$, Mann-Whitney). Importantly, comparing the distributions of winning rates across stimulation groups, I found that the reduction in cheating rates resulting from anodal tDCS was most pronounced in participants who experienced a high degree of conflict between moral and selfish motives, i.e., those who lied only on occasion (incomplete cheaters) and not always (complete cheaters): while the number of incomplete cheaters was drastically reduced in the anodal stimulation group, complete cheaters remained roughly the same irrespective of the experimental condition. These findings are consistent with the hypothesis that the rDLPFC is arbitrating between conflicting motives. To further test this

hypothesis I separately analyzed cheating rates in participants who reported low or high moral conflict associated with cheating. This analysis revealed a significant difference in cheating rates between the anodal and sham groups only for high-conflict participants (p = 0.014, rank-sum test, n=54) and not for low-conflict participants (p = 0.327, rank-sum test, n = 42). Finally, I also did not find any tDCS-induced changes in behavior for measures of altruism, impulsivity, risk and ambiguity aversion, and of civic cooperation, suggesting that the effect of anodal tDCS was specific for increasing honesty. Taken together, these findings allow me to identify a specific function of the right DLPFC, namely that of resolving a decision-conflict resulting specifically between a moral motive and a material one.

Conclusion

My results support the conclusion that honesty has a biological basis in neural activity patterns in the right DLPFC. In particular, the identified patterns seem to play an important role in mediating between two conflicting motives, i.e. a self-interest motive to maximize earnings and a moral motive to conform to the behavior prescribed by the honesty norm. Further, given the selectivity of tDCS effects for the honesty task, it seems that these neural processes do not functionally overlap with other mechanisms for behavioral control related to risk (Kuhnen & Knutson 2005), timing of rewards (Kable & Glimcher 2007), or altruism (Quervain et al. 2004).

2.3. Study 3: Causal neural networks underlying social norm compliance

Introduction

Several studies have highlighted the vital importance of institutionalized sanction threats for the maintenance of social order across most human societies. Previous research proposed that

the human brain has developed distinct neural mechanisms that mediate norm-compliant social behavior in response to such punishment threats (Fehr & Gächter 2002). In particular, fMRI studies (Spitzer et al. 2007) identified the neural networks activated during voluntary social norm compliance as well as those responding to the introduction of credible punishment threats for potential norm violators (i.e. sanction-induced norm compliance). Moreover, a recent tDCS study (Ruff et al. 2013) demonstrated a causal functional role of one of the regions central to this identified network, i.e. the right lateral prefrontal cortex (LPFC), in driving norm compliant behavior for both sanction-induced and voluntary norm compliance. This study showed that norm compliance was affected in opposite ways depending on the type of stimulation (anodal vs. cathodal) and on the presence or absence of sanction threats. Notably, the effects on norm-compliance induced by cathodal tDCS were replicated in an independent study that used a different brain-stimulation method, i.e. TMS, to disrupt brain activity in the same brain region (Strang et al. 2015).

To date, however, no study has investigated at the level of neural networks which neural mechanisms may be modulated by right LPFC activity in order to implement voluntary and punishment-induced norm-compliance respectively. At present it is unclear whether only local rLPFC neural processes may be responsible for generate the norm-compliant decisions or whether rLPFC is coordinating the activity in other interconnected brain networks that are jointly responsible for the change of norm-compliant behavior. Answering this question is fundamental for achieving a more detailed mechanistic understating of how social emotional and cognitive mechanisms interact in order to determine moral behavior. Here I aim to generate the missing evidence using the concurrent combination of online tDCS and fMRI to investigate the dynamic changes in functional interplay between the stimulated right LPFC and interconnected brain areas underlying voluntary as well as sanction-induced social norm compliance. The prefrontal cortex, and the

right LPFC in particular, has been suggested to have a critical role not only in regulating activity in interconnected brain regions, but also in receiving and integrating inputs from numerous connected regions such as the amygdala, the mPFC or the parietal cortex in order to instantiate behavioral control and action selection of complex processes (Duncan 2001; Miller & Cohen 2001; Buckholtz et al. 2015). One can therefore expect that modulation of the right LPFC leads to changes in brain regions such as the ACC or the STS responsible for detecting when violating social norms is relatively safe and dynamically adapting behavior where the possibility to be punished is absent. Furthermore, it is also plausible to expect that tDCS induced changes in this region's excitability could result in different level of neural activity in regions that are responsible for processing emotional responses to social punishment threats such as the amygdala, the insula or the OFC.

Comparing this newly obtained data with the existing evidence could disclose if the network of regions involved in norm compliance overlaps with regions previously implicated in moral-specific processes, such as the OFC activity associated with processing morally salient emotions (Heekeren et al. 2005; Moll et al. 2002). It is important to note, however, that due to the absence of a direct comparison, the overlaps potentially observed cannot be taken as conclusive evidence, but only as insightful hints for domain-specificity that would need to be directly tested in future research.

Methods

Seventy-nine healthy female volunteers (mean age, 21.56 years; SD, 5.05 years) were randomly assigned to one of the three stimulation groups: anodal, cathodal, and sham tDCS. Participants completed in individual sessions the exact same monetary allocation task used by Spitzer et al. (2006) and Ruff et al. (2013) while undergoing fMRI and concurrently receiving one of the three types of tDCS. Briefly, the task completed by the participants consisted of

anonymously dividing a given monetary amount between themselves and another person: on every trial, player A ("the proposer") randomly interacted with a player B ("the responder"). At the beginning of each trial, player A was endowed with 100 money units (MUs) and proposed a division of these MUs between himself and player B. Both players also received 25 MUs extra which could not be transferred.

In the control condition (no punishment condition), the monetary transfer was implemented exactly as proposed by Player A decided, with no intervention allowed for Player B. In contrast, in the punishment condition, player B could use any amount of the extra 25 MUs to punish player A with the following costs scheme: for every MU spent by player B for punishment lead to a reduction of player A's gain by 5 MUs. Inside the scanner, all participants played the role of player A, "the proposer". In total, each player A underwent 90 trials, 45 under the threat of punishment and 45 control trials.

Brain stimulation was delivered using two 5x7 cm MR-compatible stimulation electrodes, the active one placed over the right LPFC (active electrode at: x = 52, y = 28, z = 14 MNI coordinates; note these are the same coordinates used in the previously mentioned tDCS study (Ruff et al. 2013)), and the reference electrode placed over the vertex. Importantly the brain stimulation protocol mirrored closely the protocol used by (Ruff et al. 2013) with respect to electrode montage and stimulation intensity (1 mA). The only change was in the duration of the active stimulation, lasting for 30 minutes for anodal and cathodal stimulations (for the sham condition the stimulation lasted 30 seconds). This change was necessary due to the extended duration of the task, which entailed more trials than in the previous behavioral tDCS study (Ruff et al. 2013) to ensure sufficient statistical power for the fMRI analyses.

Results

Critically, at the behavioral level, the results of the present study replicated previous findings obtained in a tDCS study employing the very same task (Ruff et al. 2013): anodal and cathodal tDCS changed sanction-induced norm compliance in opposite ways relative to sham tDCS. Specifically anodal tDCS increased the monetary transfer difference between punishment and control trials (LME analysis, $p < 0.001$), whereas cathodal tDCS decreased this monetary transfer difference (LME analysis, $p = 0.001$).

At the neural level, I identified the brain network sub-serving social-norm-compliant behavior with a contrast comparing decision-making in trials entailing a social punishment threat and the decisions in the control trials in the baseline group (i.e. receiving sham stimulation). In line with previous neuroimaging evidence, (Spitzer et al. 2007), I found increased brain activity in response to punishment threats in the so-called executive network, including the bilateral DLPFC, left VLPFC, and ACC.

More importantly, the central analysis of this paper aimed at identifying at the neural-network level the changes induced by tDCS of the right dlPFC. Critically, this analysis can reveal if there are differences in the neural networks affected by different modulations of dlPFC excitability. Our hypothesis was that the modulation of the sanction-induced norm-compliant behavior by rLPFC-tDCS would be reflected by activity changes in two distinct networks. Here we can test if increasing rLPFC excitability with anodal tDCS or decreasing it via cathodal tDCS would have effects on similar neural networks or instead if either stimulation would selectively induce changes in different networks.

To test for these potential differential effects of upregulating and downregulating rLPFC excitability, I performed a two-sample t-tests revealing that tDCS led to differential cortical sensitivity changes in response to the punishment threat in several brain regions. In detail, the increase of monetary transfers in the presence of sanction threats following anodal stimulation was reflected at the neural level by an increase in amygdala response to the

sanction threats following the increase in neural excitability of the right dlPFC induced by anodal tDCS ($p < 0.05$) I further found functional connectivity changes, revealed by a psychophysiological interaction (PPI) analysis, revealing that anodal tDCS lead to an increase in coupling between the right dlPFC and the orbitofrontal cortex (OFC). Importantly, these results offer a perspective of the neural network responding to punishment threats at the moment of determining the decision to comply with the fairness social norm.

Conversely, the opposite behavioral pattern observed following cathodal tDCS disrupting neural excitability of the right LPFC was reflected in a decrease in functional activity in the anterior cingulate cortex (ACC), bilateral LPFC, and left parietal cortex. These results suggest that a decrease in right LPFC excitability lead to a reduction of ability of strategically switching strategies depending on the presence of punishment threats, since these brain areas have been previously associated with central executive network functions responsible for dynamically regulating behavior (Spreng et al. 2013; Bressler & Menon 2010).

Conclusion

The present study aimed at elucidating, at the level of neural networks, the changes induced by tDCS stimulation of the right LPFC that are responsible for changes in fairness compliance behavior (measured via monetary allocations). Such changes have been shown to be causally inked to neural excitability of this region by two independent studies using different brain stimulation methods: a previous study (Ruff et al. 2013) used tDCS to both increase and decrease right dlPFC excitability, while the second one used TMS (Strang et al. 2015) to decrease dlPFC activity. In the present study I critically replicated the behavioral effects of these studies, thereby consolidating the view that the right LPFC is causally involved in both voluntary norm compliance as well as enforced norm-compliance.

At the neural level, I could identify the changes in neural activity and connectivity that reflected the reorganization of the neural processes driving behavior, as a result of the neuromodulation induced by tDCS. More in detail, anodal stimulation enhanced neural sensitivity to punishment threats in the amygdala and triggered stronger punishment-related connectivity between the stimulated LPFC and the OFC, suggesting that the tDCS may have increased affective responses to punishment threats. In contrast, cathodal stimulation affected brain regions within a central executive network, consistent with the view that stimulation may have modulated strategic behavior triggered by the consequences of the punishment threats.

Crucially, these results demonstrated that increasing or decreasing LPFC excitability had drastically different effects on the neural processes regulating social norm compliance in the presence of punishment threats. Intriguingly, increasing LPFC excitability with anodal tDCS enhanced the affective responses to the presence of a punishment threat. In contrast decreasing rLPFC excitability with cathodal tDCS triggered changes in strategic responses to the presence of a punishment threat, by modulating changes in regions within the executive network.

It is plausible that following anodal tDCS the LPFC increases its receptivity to the inputs from the affective network, an interpretation in line with the proposed the integration-and-selection role of this area in the service of norm-compliant behavior (Buckholtz and Marois, 2012; Buckholtz et al., 2015): higher LPFC involvement in the decision to comply with the fairness social norm resulted in an increased awareness of the potential negative consequences of being punished for violating the fairness norm. This is consistent with other research (Pripfl et al., 2013) reporting that increased LPFC excitability lead to reduced willingness of incurring risks but only when these choices were affectively charged.

Similarly, decreasing the LPFC excitability with cathodal tDCS could have disrupted its integration-and-selection role, ultimately resulting in an inability of updating one's behavior between different choice situations – i.e. updating the monetary transfers from trials where the punishment threat was absent to the ones where it was present. Consequently, this diminished integration ability resulted in a weaker representation of the punishment threat signal in the executive control regions devoted to regulate social norm-compliant behavior. This interpretation is consistent with previous findings demonstrating that disrupting LPFC activity with TMS resulted in an inability of selectively updating one's behavior in an iterated prisoners dilemma task (Soutschek et al., 2015). These conclusions should be taken with a grain of salt since they are they rely on reverse inference speculation and should be directly tested in future experiments. Nonetheless these results are an important step forward in the understanding of how the LPFC orchestrates social norm compliance modulating the activity of specific neural networks depending of the context in which the decisions are taken.

Taken together, my findings show that the causal role of the stimulated LPFC region in regulating social norm compliance may result from a dynamic modulation of brain regions involved in affective and strategic responses to sanction threats. Finally, these findings have important behavioral implications for understanding norm-compliance and designing institutions that promote fairness. On the one hand, these results show that strategic thinking decreases the compliance with fairness norms in situations where norms are not enforced by punishment threats. On the other hand, the possibility of being punished triggers norm compliance by acting on our sensitivity to the threat of receiving social punishment. Intriguingly, I found that the same brain region regulates both our sensitivity to punishment threats as well as our ability of strategic thinking, therefore implying that one may inevitably also affect the other.

3. General Discussion

This thesis aimed to obtain evidence that the human brain evolved mechanisms that are specifically dedicated to favor the emergence and persistence of fundamental moral norms (Dunbar 2009). From an evolutionary perspective, moral-specific mechanism are theoretically plausible, since they may ultimately favor groups that possessed such mechanisms to regulate inter-group conflicts, probably thanks to better within-group cooperation (Henrich 2016). The three studies outlined here contributed to furthering the understanding of the neuro-psychological architecture of moral decision-making by a) linking subjective differences in moral decisions to neural representations of moral subjective values, providing a first glance at the potential neural origins of moral preferences, b) identifying neural processes that encode specifically moral value representations, c) providing a causal understanding of the role of the dlPFC in implementing moral behavior (e.g. being honest) in the presence of conflicting alternative material opportunities, e.g. behaving selfishly to maximize earnings, and d) identifying different neural-network decision mechanisms that underlie either fairness norm-compliant choices or unfair (selfish) decisions.

3.1.Neural Representations of Subjective Moral Values

In study 1, I approached moral decisions from a value-based perspective, investigating if I could explain moral decisions from the subjective value people assign to the different choice options. I further tested whether the common intuition that moral values – in the case of the present study the value of human lives – are evaluated in a similar currency as other objects. I did so by testing for differences in the neural valuation processes involved in these two qualitatively different domains.

The results of this study show that, similarly to economic choices, moral decisions about who should be saved/harmed can be well characterized by estimating the subjective value

that people assign to the proposed choice alternatives. Further, I found evidence that computations of moral subjective values, although partially relying on domain-general choice mechanisms, are largely represented in the brain by moral domain-specific decision processes.

More in detail, I found that moral decisions to not kill the one person were supported by neural activations in the right TPJ, the PCC and the right dlPFC. These results are in line with previous research (Majdandžić et al. 2012; Kahane et al. 2012), showing that judgements that it is morally impermissible to harm others in trolley-dilemma types of situations elicited patterns of activation that extensively overlapped with those identified in my study. These brain areas have been repeatedly associated with processing empathy (Bzdok et al. 2012) as well as harm aversion (Crockett et al. 2010) in decision-situations similar to the ones employed by my study. In the value-based framework proposed by my study, activity in these brain areas may thus encode the aversion to harming a person and feeding this negative-value into the moral-valuation process, thereby supporting the moral decisions that it is impermissible to harm a person in order to save more. This view is further supported by my evidence that neural activity in the rTPJ was specifically involved in the representation of moral subjective values compared to financial values.

On the other hand, the value of the larger group of people was identified to be coded by neural activity in the left IPL and bilateral Anterior Insula. Such a pattern of activity is in accordance with previous data that associated neural activity in these areas with moral judgments that it is morally appropriate to sacrifice the life of one in order to save the lives of more people (Kahane et al. 2012; Greene et al. 2004; Greene et al. 2001).

Further, activity in these brain areas was previously shown (Shenhav & Greene 2010) to positively correlate with an increase of expected value in moral decisions with probabilistic outcomes. Interestingly, in the present study I found that both left IPL and the right Anterior

Insula were involved in the computation of subjective values in both the moral and the financial domain, supporting the view that domain-general processes contribute to moral decisions. Taken together, these results support the view (Greene 2015) that moral preferences for saving a larger number of people are predominantly relying on cognitive neural mechanisms. My evidence seems to further suggest that these mechanisms are partially overlapping with those involved in value-calculations in non-moral decision making, as for instance in financial decisions.

3.2. LPFC and moral decision making

Having established that moral values are represented by domain-specific neural mechanisms in Study 1, the two following studies aimed at characterizing the function of a brain region thought to be responsible for negotiating conflicts between moral and non-moral (e.g. material/monetary) values. Based on a large body of literature, I hypothesized that a plausible candidate region for this function could be the right LPFC. Thus, in the present thesis I discuss two studies that targeted different portions of this brain region with tDCS, in order to establish a causal link between neural activity in these brain regions and different types of moral behavior: honesty and fairness. The two studies not only provided convergent evidence for a crucial causal role of the LPFC in influencing moral behavior, but also provided evidence that this region is specifically important for regulating behavior in moral contexts. In Study 2, I found that increasing LPFC excitability by means of anodal tDCS lead to an increase of honest behavior, further showing that the role of the LPFC in this context is to arbitrate between conflicting motives – a moral motive and a selfish-one – by down-tuning the latter and hence increasing people's ability to resist the temptation of lying.

My findings align with previous evidence that right LPFC activity is elicited in decision situations where people face conflict, such as intertemporal choices where one has to arbitrate between the impulse of taking a smaller immediate reward or decide to wait in order

to receive a larger reward in the future (Kable & Glimcher 2007). Simlarly, Hare and colleagues (2009) found that neural activity in this brain area correlates with the ability to exert self-control in dietary decisions, a function that is suggested to be implemented by modulating neural activity in value-related systems, particularly in the vmPFC.

Furthermore, recent studies have highlighted the causal nature of right LPFC activity in these types of cognitive control. For instance, a decrease in the ability of delaying gratification was observed following TMS-induced disruption of LPFC activity (Figner et al. 2010); Using tDCS, a previous study (Loftus et al. 2015) showed that exertion of self-control could be enhanced by upregulating the excitability of the LPFC by means of anodal stimulation. While these studies support the hypothesis that the LPFC is involved in behavioral control, my study proposes that there may be neural populations within this area recruited selectively for different types of conflict. Critically, the results of my study revealed that the stimulated portion of the LPFC was selectively involved in arbitrating conflicts between moral vs. selfish motives, since I found no stimulation effects on inter-temporal choices or risk and ambiguity aversion. The functional specificity of this brain region for conflicts involving moral values aligns with the hypothesis that the human brain evolved moral-specific mechanisms to regulate social interactions.

My results further strengthen the view that modulating neural activity in the LPFC does not influence the beliefs, or preferences, held by a person. As in other studies (Ruff et al. 2013; Knoch & Fehr 2007) where it was shown that stimulation of the right LPFC left fairness-related beliefs unaltered, I found no group differences with respect to the beliefs on the moral wrongness of lying that participants held. Interestingly, in Study 2 I did not find an effect of brain stimulation on a dictator game used to measure the value that participants assigned to money. This result further suggests that the functional role of the portion of LPFC modulated by tDCS is distinct from the more dorsal portion targeted in Study 3, where

behavior in a structurally similar task was in fact affected by both anodal and cathodal stimulation.

Study 3 replicated previous behavioral results (Ruff et al. 2013; Strang et al. 2015) implicating the LPFC in regulating fairness norm-compliant behavior, both when fairness was enforced by the presence of a punishment threat and when compliance with the norm was voluntary: Sanction-induced norm compliance was increased by anodal tDCS and decreased by cathodal tDCS, whereas voluntary compliance was reduced by anodal tDCS and enhanced by cathodal tDCS. This pattern of results suggests that anodal/cathodal tDCS rendered participants more/less sensitive to the presence of sanction threats, causing larger/smaller adjustments of behavior to the external incentives.

The neuroimaging data obtained in study 3 found a pattern of results consistent with this view: increasing dlPFC excitability lead to a stronger involvement of a neural network believed to process threats as well as moral emotions, mainly in the OFC, which at the behavioral level resulted in an increased compliance with the fairness norm when a punishment threat was present. Conversely, decreasing dlPFC excitability resulted in a decrease in neural functioning in neural regions involved in strategic behavioral planning in response to the absence/presence of punishment threats, including the ACC, the IPL, left dlPFC, cuneus and precuneuns. These neural changes where reflected at the behavioral level in participants' transfer decisions: receiving cathodal stimulation over the right LPFC lead participants to behave less strategically, systematically transferring more when punishment threats were absent and less when these were present. This behavior in the latter condition was particularly sub-optimal since it increased the amount of punishment received by their counterparts.

Jointly, studies 2 and 3 revealed an important causal role of the right LPFC in different types of behavior. The evidence obtained in these two studies suggests that the

LPFC seems to be exerting a regulatory role, increasing or decreasing the involvement of other brain areas in the mechanisms underpinning moral behaviors.

3.3.Future Directions

The studies discussed above build on research that has yielded a sophisticated understanding of the latent principles that guide moral behaviors (Haidt & Joseph, 2004, Cushman et al., 2006, Gray & Wegner, 2011, Mikhail, 2011, Malle et al., 2014) as well as detailed maps of the neural substrates supporting these behaviors (Greene et al. 2004; Young & Dungan 2012). However, given this extraordinary literature on the psychological mechanisms supporting moral intuition, it is remarkable how little we understand about their origins. Future studies should aim to close this gap, for instance by providing evidence answering a simple and crucial question: Where do moral intuitions come from?

Recently, two influential papers (Crockett 2013; Cushman 2013) proposed that the origins of moral intuitions may be related to social learning, a mechanism that is crucial for the evolvement of preferences in value-based decision making (Clithero & Rangel 2014) and for the evolution of social norms, both in laboratory settings (Peysakhovich & Rand 2016) and in field studies in small-scale societies around the world (Henrich et al. 2001). Future studies should aim to increase understanding of how personal moral values can change in the process of learning about others' moral preferences, hence disclosing the neural and cognitive mechanisms underlying the social learning of moral values. Some important questions which could guide this research include: 1) What cognitive processes best characterize how others' moral values influence our own moral values? 2) What personal characteristics of others determine whether and the extent to which we learn their moral values? 3) What brain structures show neural activity that correlates with the evolvement of moral values? 4) Which of these brain regions and processes are causally necessary for learning moral values and integrating them into previously held moral values? The research discussed in the present

thesis provides crucial evidence illustrating how moral value computations are implemented in the brain. This represents a first crucial step that allows developing future research on how moral value computations evolve and are affected by the social context.

Evidence answering these questions would provide crucial first insights into the learning mechanisms responsible for the evolution of moral values. Combining behavioral and neuroscientific measurements in particular would allow me to focus on neural plasticity and its relation to changeability of moral preferences. Moreover, such knowledge has the potential to further the understanding of the dynamics of moral values within societies and of the design of policies and institutions that may favor dialogue between people with different moral preferences.

From a more methodological perspective, while brain stimulation studies provide a promising starting point for a more mechanistic causal understanding of the neural mechanisms steering our social behavior, these have not yet addressed a whole range of fundamental questions that may guide research in the coming years. For instance, many models of the processes contributing to social behavior have been specified at a purely conceptual level, without a quantitative formalization, making it difficult to investigate how brain stimulation affects these processes. Promising avenues in this direction therefore include combining brain stimulation methods with computational models of social learning and choice processes which more explicitly formulate distinct neural computations that may be mediated by the stimulated brain area. In the studies discussed in this thesis, I have proposed a novel paradigm to estimate subjective moral values in a mathematically precise way, and I have successfully modulated behavior with tDCS. These studies therefore provide some fundamental methods that can guide future research aiming to combine them in more complex designs.

4. General Conclusions

Moral behavior is critical for the orderly maintenance of human societies, and by increasing within group cooperation has been fundamental for surviving inter-group wars (Henrich 2016). The emergence and maintenance of moral norms has been proposed to have been favored by the evolution of emotional and cognitive neural processes dedicated specifically to regulating social interactions. In the present thesis, I investigated if we can identify some of these moral-specific processes from different perspectives. I studied the existence of mechanisms a) supporting the moral intuitions at the basis of subjective moral values, which can explain individual differences in moral preferences, b) regulating the salience of moral motives influencing behaviors when facing situations presenting selfish temptations, and c) orchestrating behavioral responses in the presence of social norms that prescribe actions.

In order to better understand the origins of moral intuitions, I relied on the literatures on value-based decision-making and on moral intuitions to investigate the hypothesis that moral preferences can be captured by mathematical models that characterize the individual subjective value representations of human lives. I found behavioral evidence supporting this hypothesis, being able to reliably estimate subjective moral values with a model traditionally used to capture financial subjective values. Having mathematically characterized the participants' moral value functions, I could further detect neural signals underlying different moral preferences, linking more utilitarian individuals to neural activity in brain areas linked to cognitive reasoning, such as the left IPL (Shenhav & Greene 2010; Greene et al. 2004), while individuals with more non-consequential moral preferences revealed stronger activations in a network of regions related to empathy and harm aversion, such as the right TPJ and anterior insula (Crockett et al. 2014; Majdandžić et al. 2012; Bzdok et al. 2012).

Further, I used tDCS to address the question of whether the right LPFC plays a crucial causal role in modulating moral behavior. To this end, I combined this brain stimulation method with both behavioral measures of moral choice and neuroimaging, providing several

pieces of evidence for a direct causal involvement of this region in regulating moral behavior. Moreover, in Study 3, I could further show that the modulation by tDCS of the neural networks involved in social norms compliance. Taken together, the evidence discussed in the present thesis supports the hypothesis that the LPFC has a regulatory role that determines the extent to which different neural mechanisms (which respond to different aspects of a given decision situation) regulate behavior.

My results revealed that increasing LPFC brain activity resulted in increased capacity of resisting the temptation of lying in order to achieve a selfish reward, in order to instead follow a moral motive of being honest. Further, LPFC activity regulates behavior by modulating neural functioning in the neural network underpinning norm compliance. I indeed found that increased LPFC activity resulted in increasing responsiveness to punishment threats in order to avoid punishment while decreased excitability in the LPFC resulted in decreased activity in the central executive network, which at the behavioral level was reflected by a putative reduction in the ability to strategically decisions to the different choice situations.

In conclusion, in the present thesis I gathered evidence collectively supporting the Social Brain Hypothesis, and in the process helped shedding light on the neuro-psychological mechanisms involved in moral behavior. More generally, these findings may contribute to society by informing policy making and institutional design aimed at preventing antisocial and criminal behavior, as well as diagnosis and treatment of pathologies related to impaired moral decision-making, such as psychopathy (Cima et al. 2010; Moll et al. 2002).

References

Abe, N. et al., 2014. The neural basis of dishonest decisions that serve to harm or help the target. *Brain and Cognition*, 90, pp.41–49. Available at: http://dx.doi.org/10.1016/j.bandc.2014.06.005.

Apperly, I.A., 2008. Beyond Simulation-Theory and Theory-Theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind." *Cognition*, 107(1), pp.266–283.

Axelrod, R., 1984. *The evolution of cooperation*, Available at: http://www.sciencemag.org/content/211/4489/1390.short.

Ayars, A., 2016. Can model-free reinforcement learning explain deontological moral judgments? *Cognition*, 150, pp.232–242. Available at: http://dx.doi.org/10.1016/j.cognition.2016.02.002.

Barton, R. & Dunbar, R., 1997. Evolution of the social brain. *Machiavellian Intelligence II: Extensions and Evaluations*, (November), p.403. Available at: http://books.google.com/books?hl=en&lr=&id=bV9yeFV6_ckC&pgis=1.

Batson, C.D., 2011. What's wrong with morality? *Emotion Review*, 3(3), pp.230–236. Available at: http://emr.sagepub.com/cgi/doi/10.1177/1754073911402380.

Baumgartner, T. et al., 2011. Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nature Neuroscience*, 14(11), pp.1468–1474. Available at: http://dx.doi.org/10.1038/nn.2933.

Berkowitz, L., 2003. Affect, aggression, and antisocial behavior. In *Handbook of affective sciences*. pp. 804–823.

Bicchieri, C., 2005. *The Grammar of Society*,

Blair, R.J.R., 2007. Empathic dysfunction in psychopathic individuals. *Empathy in mental illness*, (September), pp.3–16. Available at: http://www.hum.utah.edu/~bbenham/Minds and Morals/06 Blair Empathy in psychopathy Chapter in Empathy in psychiatric illness.pdf.

Blair, R.J.R., 2007. The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences*, 11(9), pp.387–392.

Botvinick, M.M. et al., 2001. Conflict monitoring and cognitive control. *Psychological Review*, 108(3), pp.624–652. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.108.3.624.

Bressler, S.L. & Menon, V., 2010. Large-scale brain networks in cognition: emerging methods and principles. *Trends in Cognitive Sciences*, 14(6), pp.277–290.

Buckholtz, J.W. et al., 2015. From Blame to Punishment: Disrupting Prefrontal Cortex Activity Reveals Norm Enforcement Mechanisms. *Neuron*, 87(6), pp.1369–1380.

Bzdok, D. et al., 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure and Function*, 217(4), pp.783–796. Available at: http://link.springer.com/10.1007/s00429-012-0380-y.

Cima, M., Tonnaer, F. & Hauser, M.D., 2010. Psychopaths know right from wrong but don ' t care. , (2009).

Clithero, J.A. & Rangel, A., 2014. Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, 9(9), pp.1289–1302. Available at: http://scan.oxfordjournals.org/lookup/doi/10.1093/scan/nst106.

Cohen, J.D. et al., 1997. Temporal dynamics of brain activation during a working memory task. *Nature*, 386(6625), pp.604–608.

Cohen, J.D., Dunbar, K. & McClelland, J.L., 1990. On the control of automatic processes: a

parallel distributed processing account of the Stroop effect. *Psychological review*, 97(3), pp.332–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2200075.

Crockett, M.J. et al., 2014. Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences of the United States of America*, 111(48), pp.17320–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4260587&tool=pmcentrez&rendertype=abstract.

Crockett, M.J., 2016. How Formal Models Can Illuminate Mechanisms of Moral Judgment and Decision Making. *Current Directions in Psychological Science*, 25(2), pp.85–90. Available at: http://cdp.sagepub.com/content/25/2/85.abstract.

Crockett, M.J., 2013. Models of morality. *Trends in Cognitive Sciences*, 17(8), pp.363–366.

Crockett, M.J. et al., 2010. Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences*, 107(40), pp.17433–17438. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1009396107.

Cushman, F., 2013. Action, outcome, and value: a dual-system framework for morality. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc*, 17(3), pp.273–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23861355.

Cushman, F. & Young, L., 2011. Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, 35(6), pp.1052–1075.

Cushman, F., Young, L. & Greene, J.D., 2010. Multi-system Moral Psychology. *The Moral Psychology Handbook*, pp.1–20.

Cushman, F., Young, L. & Hauser, M., 2006. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, 17(12),

pp.1082–1089.

Dávid-Barrett, T. & Dunbar, R.I.M., 2013. Processing power limits social group size: computational evidence for the cognitive costs of sociality. *Proceedings. Biological sciences / The Royal Society*, 280(1765), p.20131151. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23804623.

Dresler, T. et al., 2009. Emotional Stroop task: Effect of word arousal and subject anxiety on emotional interference. *Psychological Research*, 73(3), pp.364–371.

Dunbar, R.I.M., 1998. The Social Brain Hypothesis. *Evolutionary Anthropology*, pp.178–190.

Dunbar, R.I.M., 2009. The social brain hypothesis and its implications for social evolution. *Annals of Human Biology*, 36(5), pp.562–572.

Dunbar, R.I.M. & Shultz, S., 2007. Evolution in the social brain. *Science (New York, N.Y.)*, 317(5843), pp.1344–1347. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17823343%5Cnhttp://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&DbFrom=pubmed&Cmd=Link&LinkName=pubmed_pubmed&LinkReadableName=Related Articles&IdsFromResult=17823343&ordinalpos=3&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed.

Duncan, J., 2001. An adaptive coding model of neural function in prefrontal cortex. *Nat Rev Neurosci*, 2(November), pp.820–829. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11715058.

Elster, J., 1989. Social Norms and Economic Theory. *Journal of Economic Perspectives*, 3(4), pp.99–117. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=4432933&site=ehost-live.

Fadiga, L., Craighero, L. & Ausilio, A.D., 2010. Chapter 14 Broca ' s area in language , action , and music. *Annals Of The New York Academy Of Sciences*, 1169(1), pp.448–458. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.2009.04582.x/full.

Fadiga, L., Craighero, L. & Roy, A., 2009. Broca's Region: A Speech Area? In *Broca's Region*.

Fehr, E. & Gächter, S., 2002. Altruistic punishment in humans. *Nature*, 415(6868), pp.137–140. Available at: http://www.nature.com/doifinder/10.1038/415137a.

Figner, B. et al., 2010. Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience*, 13(5), pp.538–539. Available at: http://www.nature.com/doifinder/10.1038/nn.2516.

Foot, P., 1967. The Problem of Abortion and the Doctrine of the Double Effect. *OxFoot, P., 1967. The Problem of Abortion and the Doctrine of the Double Effect. Oxford Review, (5), pp.5–15.ford Review*, (5), pp.5–15.

Gazzaniga, M.S., 2002. *Experimental techniques used with animals*, Available at: http://books.google.com/books?id=CdNqlAEACAAJ&dq=gazzaniga+cognitive+neuroscience+(biology+of+the+mind+norton)&hl=&cd=2&source=gbs_api%5Cnpapers3://publication/uuid/40BDFB99-B5E0-4CBC-88E8-7F44D2550D8C.

Gleichgerrcht, E. & Young, L., 2013. Low Levels of Empathic Concern Predict Utilitarian Moral Judgment. *PLoS ONE*, 8(4).

Gordon, R.M., 1996. "Radical" simulationism. In *Theories of Theories of Mind*. pp. 11–21. Available at: file:///Users/Domingo/Documents/Literature/E_Lit/.

Gray, K. & Wegner, D.M., 2009. Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of personality and social psychology*, 96(3), pp.505–520.

Green, L. et al., 2004. Discounting of delayed food rewards in pigeons and rats: is there a

magnitude effect? *Journal of the experimental analysis of behavior*, 81(1), pp.39–50. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1284970&tool=pmcentrez&rendertype=abstract.

Greene, J. & Haidt, J., 2002. How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12), pp.517–523. Available at: http://www.sciencedirect.com/science/article/pii/S1364661302020119.

Greene, J.D., 2015. *Moral tribes : emotion, reason, and the gap between us and them*,

Greene, J.D. et al., 2004. The Neural Bases of Cognitive Conflict and Control in Moral Judgment. , 44, pp.389–400.

Greene, J.D., 2007. Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8), pp.322–323.

Greene, J.D. & Paxton, J.M., 2009. Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, 106(30), pp.12506–12511.

Greene, J.D., Sommerville, R.B. & Nystrom, L.E., 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. , 293(September), pp.2105–2108.

Güth, W., 1995. On ultimatum bargaining experiments - A personal review. *Journal of Economic Behavior and Organization*, 27(3), pp.329–344.

Güth, W., Schmittberger, R. & Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3(4), pp.367–388.

Haidt, J., 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological review*, 108(1), pp.814–834.

Haidt, J., 2007. The new synthesis in moral psychology. *Science (New York, N.Y.)*, 316(5827), pp.998–1002.

Haidt, J., 2012. The righteous mind. *Why Good People are Divided by Politics and Religion ...*, (January), pp.1–508. Available at: http://www.sce.cornell.edu/sce/altschuler/pdf/altschuler_review_20120301_460.pdf%5Cnpapers3://publication/uuid/AE91E6CE-E7BE-4F35-85A6-1007295862C3.

Haidt, J. & Joseph, C., 2004. Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus*, 133, pp.55–66.

Hamilton, D.L., 2005. Social cognition : key readings. *Key readings in social psychology*. Available at: http://www.loc.gov/catdir/enhancements/fy0647/2004009845-d.html%5Cnhttp://firstsearch.oclc.org/WebZ/DCARead?standardNoType=1&standardNo=086377590X:srcdbname=worldcat:fromExternal=true&sessionid=0.

Hare, T.A., Camerer, C.F. & Rangel, A., 2009. Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. *Science*, 324(5927), pp.646–648.

Hare, T. a et al., 2008. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(22), pp.5623–5630.

Harlé, K.M. & Sanfey, A.G., 2010. Effects of approach and withdrawal motivation on interactive economic decisions. *Cognition & Emotion*, 24(8), pp.1456–1465. Available at: http://www.tandfonline.com/doi/abs/10.1080/02699930903510220.

Hauser, M., 2007. Moral Minds: The Nature of Right and Wrong. *Psychology*, p.530.

Heekeren, H.R. et al., 2005. Influence of bodily harm on neural correlates of semantic and moral decision-making. *Neuroimage*, 30(313–324), pp.887–97. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15652323.

Henrich, J. et al., 2001. In search of Homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2), pp.73–84.

Henrich, J., 2016. The secret of our success: How culture is driving human evolution,

domesticating our species, and making us smarter. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Available at: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc12&NEWS=N&AN=2016-18797-000.

Huebner, B., Dwyer, S. & Hauser, M., 2008. The role of emotion in moral psychology. , (December).

Hume, D., 2000. A Treatise of Human Nature. *A Treatise of Human Nature*, 26(1739), p.626.

Hutcherson, C.A. et al., 2015. Emotional and Utilitarian Appraisals of Moral Dilemmas Are Encoded in Separate Areas and Integrated in Ventromedial Prefrontal Cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35(36), pp.12593–605. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26354924%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4563040.

Jensen, A.R. & Rohwer, W.D., 1966. The Stroop color-word test: a review. *Acta psychologica*, 25(1), pp.36–93. Available at: http://www.ncbi.nlm.nih.gov/pubmed/5328883.

Kable, J.W. & Glimcher, P.W., 2007. The neural correlates of subjective value during intertemporal choice. *Nature neuroscience*, 10(12), pp.1625–33. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2845395&tool=pmcentrez&rendertype=abstract.

Kahane, G. et al., 2012. The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience*, 7(4), pp.393–402. Available at: http://scan.oxfordjournals.org/lookup/doi/10.1093/scan/nsr005.

Kanwisher, 2006. Functional specificity in the human brain: A window H into the functional

architecture of the mind. *South East Asia Research*, 14(3), pp.445–469.

Kanwisher, N. & Yovel, G., 2006. The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 361(1476), pp.2109–28.

Kliemann, D. et al., 2008. The influence of prior record on moral judgment. *Neuropsychologia*, 46(12), pp.2949–2957. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18606175.

Knoblauch, K., 2002. Color Vision. *Stevens Handbook of Experimental Psychology Sensation and Perception*, 1(3), pp.41–75. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6977310.

Knoch, D. et al., 2010. A Neural Marker of Costly Punishment Behavior. *Psychological Science*, 21(3), pp.337–342. Available at: http://pss.sagepub.com/lookup/doi/10.1177/0956797609360750.

Knoch, D. et al., 2006. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, 314(5800), pp.829–832.

Knoch, D. et al., 2009. Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(49), pp.20895–20899. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2785725&tool=pmcentrez&rendertype=abstract.

Knoch, D. et al., 2008. Studying the neurobiology of social interaction with transcranial direct current stimulation - The example of punishing unfairness. *Cerebral Cortex*, 18(9), pp.1987–1990.

Knoch, D. & Fehr, E., 2007. Resisting the power of temptations: The right prefrontal cortex and self-control. In *Annals of the New York Academy of Sciences*. pp. 123–134.

Knutson, B. et al., 2005. Distributed Neural Representation of Expected Value. *The Journal of Neuroscience*, 25(19), pp.4806–4812. Available at: http://www.jneurosci.org/content/25/19/4806%5Cnhttp://www.jneurosci.org/content/25/19/4806.full.pdf%5Cnhttp://www.jneurosci.org/content/25/19/4806.long%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/15888656.

Knutson, B. & Peterson, R., 2005. Neurally reconstructing expected utility. *Games and Economic Behavior*, 52(2), pp.305–315.

Kohlberg, L., 1976. Moral stages and moralization: The cognitive-developmental approach. *Moral development and behavior: Theory, research and social issues*, pp.31–53.

Kohlberg, L., 1971. Stages of Moral Development. *Moral education*, pp.23–92.

Kuhnen, C.M. & Knutson, B., 2005. The neural basis of financial risk taking. *Neuron*, 47(5), pp.763–770.

Loftus, A.M. et al., 2015. The impact of transcranial direct current stimulation on inhibitory control in young adults. *Brain and Behavior*, 5(5).

Luhmann, C.C., 2009. Temporal decision-making: insights from cognitive neuroscience. *Frontiers in behavioral neuroscience*, 3(October), p.39.

van Maanen, L., van Rijn, H. & Borst, J.P., 2009. Stroop and picture-word interference are two sides of the same coin. *Psychonomic Bulletin and Review*, 16(6), pp.987–999. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19966248.

MacLeod, C.M., 1991. Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, 109(2), pp.163–203. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.211.5613&rep=rep1&type=pdf.

Majdandžić, J. et al., 2012. The Human Factor: Behavioral and Neural Correlates of Humanized Perception in Moral Decision Making. *PLoS ONE*, 7(10).

Malle, B.F., Guglielmo, S. & Monroe, A.E., 2014. A Theory of Blame. *Psychological Inquiry*, 25(2), pp.147–186. Available at: http://www.tandfonline.com/doi/abs/10.1080/1047840X.2014.877340.

McClure, S.M. et al., 2007. Time Discounting for Primary Rewards. *Journal of Neuroscience*, 27(21), pp.5796–5804. Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4246-06.2007.

McKenna, F.P. & Sharma, D., 1995. Intrusive cognitions: An investigation of the emotional Stroop task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(6), pp.1595–1607.

McKenna, F.P. & Sharma, D., 2004. Reversing the emotional Stroop effect reveals that it is not what it seems: the role of fast and slow components. *Journal of experimental psychology. Learning, memory, and cognition*, 30(2), pp.382–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14979812.

Mikhail, J., 2009. Moral Grammar and Intuitive Jurisprudence : A Formal Model of Unconscious Moral and Legal Knowledge. , 50(8).

Miller, E.K. & Cohen, J.D., 2001. An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, 24, pp.167–170.

Moll, J. et al., 2005. The neural basis of human moral cognition. *Nature Reviews*, 6, pp.799–809.

Moll, J. et al., 2002. The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *The Journal of Neuroscience*, 22(7), pp.2730–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11923438.

Moll, J. & de Oliveira-Souza, R., 2007. Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences*, 11(8), pp.319–321.

Moll, J., De Oliveira-Souza, R. & Zahn, R., 2008. The neural basis of moral cognition: Sentiments, concepts, and values. *Annals of the New York Academy of Sciences*, 1124, pp.161–180.

Montague, P.R. & Lohrenz, T., 2007. To Detect and Correct: Norm Violations and Their Enforcement. *Neuron*, 56(1), pp.14–18.

Ochsner, K.N. & Gross, J.J., 2005. The cognitive control of emotion. *Trends in Cognitive Sciences*, 9(5), pp.242–249.

Pascual, L., Rodrigues, P. & Gallardo-Pujol, D., 2013. How does morality work in the brain? A functional and structural perspective of moral behavior. *Frontiers in Integrative Neuroscience*, 7(September), pp.1–8. Available at: http://journal.frontiersin.org/article/10.3389/fnint.2013.00065/abstract.

Patil, I. & Silani, G., 2014. Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology*, 5(MAY).

Peysakhovich, A. & Rand, D.G., 2016. Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory. *MANAGEMENT SCIENCE*, 62(3), pp.631–647. Available at: http://dx.doi.org/10.1287/mnsc.2015.2168.

Piaget, J., 1932. The moral judgement of the child. *Penguin education*, p.399p. Available at: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0029252407.

Platt, M.L. & Huettel, S.A., 2008. Risky business: the neuroeconomics of decision making under uncertainty. *Nature neuroscience*, 11(4), pp.398–403. Available at: http://dx.doi.org/10.1038/nn2062.

Prinz, J., 2006. The emotional basis of moral judgments. *Philosophical Explorations*, 9(1), pp.29–43.

Quervain, D.J. De et al., 2004. The Neural Basis of Altruistic Punishment. *World*, 305(5688),

pp.1–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15333831.

Reeck, C., Ames, D.R. & Ochsner, K.N., 2016. The Social Regulation of Emotion: An Integrative, Cross-Disciplinary Model. *Trends in Cognitive Sciences*, 20(1), pp.47–63.

Ruff, C., Ugazio, G. & Fehr, E., 2013. Changing Social Norm Compliance with Noninvasive Brain Stimulation. *Science*, 342(6157), pp.482–484.

Ruff, C.C., Ugazio, G. & Fehr, E., 2013. Changing Social Norm Compliance With Noninvasive Brain Stimulation. *Science (New York, N.Y.)*, 415(6868), pp.137–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11805825%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/24091703.

Samson, D. & Apperly, I.A., 2010. There is more to mind reading than having theory of mind concepts: New directions in theory of mind research. *Infant and Child Development*, 19(5), pp.443–454.

Sanfey, A.G., 2007. Decision neuroscience: New directions in studies of judgment and decision making. *Current Directions in Psychological Science*, 16(3), pp.151–155.

Saxe, R., 2009. Theory of Mind (Neural Basis). In *Encyclopedia of Consciousness*.

Saxe, R., Carey, S. & Kanwisher, N., 2004. Understanding other minds: linking developmental psychology and functional neuroimaging. *Annual review of psychology*, 55, pp.87–124. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14744211.

Schnall, S., Haidt, J., et al., 2008. Disgust as Embodied Moral Judgment.

Schnall, S., Benton, J. & Harvey, S., 2008. With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science*, 19(12), pp.1219–1222.

Schultz, W., 2006. Behavioral theories and the neurophysiology of reward. *Annual review of psychology*, 57, pp.87–115. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16318590.

Shenhav, A., Botvinick, M.M. & Cohen, J.D., 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, 79(2), pp.217–40. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3767969&tool=pmcentrez&rendertype=abstract.

Shenhav, A. & Greene, J.D., 2010. Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude. *Neuron*, 67(4), pp.667–677. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0896627310005817.

Silani, G. et al., 2013. Right Supramarginal Gyrus Is Crucial to Overcome Emotional Egocentricity Bias in Social Judgments. *Journal of Neuroscience*, 33(39), pp.15466–15476. Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1488-13.2013.

Sip, K.E. et al., 2008. Detecting deception: the scope and limits. *Trends in Cognitive Sciences*, 12(2), pp.48–53.

Smith, E.E. & Jonides, J., 1998. Neuroimaging analyses of human working memory. *Proceedings of the National Academy of Sciences of the United States of America*, 95(20), pp.12061–12068.

Smith, E.E. & Jonides, J., 1997. Working memory: a view from neuroimaging. *Cognitive psychology*, 33, pp.5–42.

Spitzer, M. et al., 2007. The Neural Signature of Social Norm Compliance. *Neuron*, 56(1), pp.185–196. Available at: http://linkinghub.elsevier.com/retrieve/pii/S089662730700709X.

Spreng, R.N. et al., 2013. Intrinsic Architecture Underlying the Relations among the Default, Dorsal Attention, and Frontoparietal Control Networks of the Human Brain. *Journal of*

Cognitive Neuroscience, 25(1), pp.74–86. Available at: http://www.mitpressjournals.org/doi/abs/10.1162/jocn_a_00281.

Strang, S. et al., 2015. Be nice if you have to--the neurobiological roots of strategic fairness. *Social cognitive and affective neuroscience*, 10(6), pp.790–6. Available at: http://scan.oxfordjournals.org/content/10/6/790.abstract.

Stroop, J.R., 1935. Stroop color word test. *J. Exp. Physiol.*, (18), pp.643–662.

Thomson, J.J., 1976. Killing, letting die, and the trolley problem. *The Monist*, 59(2), pp.204–217.

Thomson, J.J., 2008. Turning the trolley. *Philosophy and Public Affairs*, 36(4), pp.359–374.

Tomasello, M., 2011. Human culture in evolutionary perspective. *Advances in culture and psychology*, pp.5–52. Available at: http://books.google.com/books?hl=en&lr=&id=WlMRXCvCXvkC&oi=fnd&pg=PA5&dq=Human+Culture+in+Evolutionary+Perspective&ots=oDQnm0fzBH&sig=zgAdfEcms3Vq9WYJLX55HZOjPKY.

Ugazio, G., Lamm, C. & Singer, T., 2012. The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion*, 12(3), pp.579–590.

Ugazio, G., Majdandžić, J. & Lamm, C., 2014. Are Empathy and Morality Linked? Insights from Moral Psychology, Social and Decision Neuroscience, and Philosophy. *Empathy in Morality*, pp.155–171.

Valdesolo, P. & Desteno, D., 2006. Manipulations of emotional context shape moral judgment. *Psychological Science*, 17(6), pp.476–477.

Wheatley, T. & Haidt, J., 2005. Hypnotic disgust makes moral judgments more severe. *Psychological Science*, 16(10), pp.780–784.

Williams, J.M., Mathews, a & MacLeod, C., 1996. The emotional Stroop task and psychopathology. *Psychological bulletin*, 120(1), pp.3–24. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/8711015.

Yoder, K.J. & Decety, J., 2014. The Good, the bad, and the just: justice sensitivity predicts neural response during moral evaluation of actions performed by others. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34(12), pp.4161–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24647937.

Young, L. et al., 2010. Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, 107(15), pp.6753–6758. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.0914826107.

Young, L. et al., 2007. The neural basis of the interaction between theory of mind and moral judgment. , 104(20).

Young, L. & Dungan, J., 2012. Where in the brain is morality? Everywhere and maybe nowhere. *Social Neuroscience*, 7(April 2015), pp.1–10.

Young, L. & Saxe, R., 2008. The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, 40(4), pp.1912–1920.

Appendix

A. Appendix to Study 1

Neural Computations Underlying Moral Preferences

Giuseppe Ugazio[*1,2], Marcus Grueschow[*1], Rafael Polania[1], Claus Lamm[3], Philippe N. Tobler[1], and Christian C. Ruff[1]

A version of this paper has been submitted to *Neuron*

[*] These authors contributed equally to this paper.

[1] Laboratory for Social and Neural Systems Research, Department of Economics,
University of Zurich, Zurich, Switzerland

[2] Moral Psychology Research Lab, Department of Psychology, Harvard University Cambridge, USA.

[3] Social, Cognitive and Affective Neuroscience Unit, Department of Psychology,
University of Vienna, Vienna, Austria

Correspondence to:

- Giuseppe Ugazio, Department of Psychology, Harvard University, 33 Kirkland St. Cambridge MA, 02138, USA, Tel: +1 857 241 73141, e-mail: giuseppe_ugazio@fas.harvard.edu.

RUNNING TITLE

Neural representations of subjective moral values

SUMMARY

Moral preferences pervade many aspects of our lives, dictating how we behave, whom we can marry, and even what we eat. Despite their relevance, one fundamental question remains unanswered: Where do individual moral preferences come from and how are they represented in the brain? It is often thought that all types of preferences reflect properties of domain-general neural decision mechanisms that employ a common “neural currency” to value choice options in many different contexts. This assumption, however, appears at odds with the observation that many humans consider it intuitively wrong to employ the same scale to compare moral value (e.g., of a human life) with material value (e.g., of money). In this paper, we directly challenge the common-currency hypothesis by comparing the neural mechanisms that represent moral and financial subjective values. In a study combining fMRI with a novel behavioral paradigm, we identify neural representations of the subjective values of human lives or financial payoffs by means of structurally identical computational models. Fitting isomorphic model variables from both domains to brain activity reveals specific patterns of neural activity that selectively represent values in the moral (in the rTPJ) or financial (in the vmPFC) domain. Thus, our findings show that human lives and money are not valued in a common neural currency, supporting theoretical proposals that human moral behavior can differ from behavior that is driven by personal material benefit.

INTRODUCTION

Moral preferences play a crucial role in determining how a person perceives the world, how she acts, and what she likes. Differences in moral preferences lie at the heart of conflicts that can ultimately lead to confrontations between individuals or even wars between nations (Berns & Atran 2012; Barnett & Pears 1997). Given their relevance, it is remarkable how little we understand about the neural and cognitive mechanisms that determine our moral preferences and that thereby underlie individual differences in our moral behavior.

In other choice domains, for example economic decisions, individual preferences have been intensely studied in terms of neural processes that assign values to choice options (Schultz 2006). Importantly, several studies have demonstrated that a person's economic preferences are reflected in subjective values encoded by activity of the ventral-medial prefrontal cortex (vmPFC), Ventral Striatum (VS), and posterior cingulate cortex (PCC, Grueschow et al. 2015; Kable & Glimcher 2007; Hare et al. 2008; Clithero & Rangel 2014; Ruff & Fehr 2014; McNamee et al. 2013; Chib et al. 2009). Based on these findings, the predominant view of human value-based decision-making posits that the brain values choice-options on a common scale that may allow us to compare and choose efficiently across many different types of goods. This has been proposed to hold not only for material goods (e.g., art, food or money) but also for non-material values (e.g., beauty, praise, or status, Levy & Glimcher 2012; Izuma et al. 2008; Zink et al. 2008). As for moral choices, recent studies (C. A. Hutcherson et al. 2015; Shenhav & Greene 2010; Crockett et al. 2017) proposed that even the value of human lives or of human pain may be computed by the same neural mechanisms that are involved in computing value of non-moral goods. However, these studies have not provided a direct link between individually determined moral preferences and neural mechanisms of value-based decision-making.

For instance, one study (Shenhav & Greene 2010) investigated neural representations of *expected values* associated with possible losses of lives (i.e., calculated in an objective fashion as the probability of death multiplied by the number of possible deaths). Such pre-defined expected-value computation is identical across different agents and therefore cannot reveal a given individual´s *subjective* valuation of the different choice options. Another study showed differences in the neural correlates of emotional and utilitarian appraisals during moral decisions (C. A. Hutcherson et al. 2015), but these findings do not reveal if differences in moral

values result from different sensitivity to these attributes, nor if individuals assigning different weights to these attributes would take different moral decisions. Finally, recent studies (Crockett et al. 2017; Crockett et al. 2014) showed that neural value responses were differentially modulated during choices about financial rewards that were coupled with painful shocks to either others or oneself. However, in this context it is impossible to know whether these neural activations indeed reflect moral concerns rather than differences in the representation of others' versus one's own affective states during pain (Silani et al. 2013; Lamm et al. 2007). Moreover, since the decisions always entailed trade-offs between pain and monetary profit, the observed neural responses in the value system still reflected the valuation of material goods. The neural processes underlying subjective valuation of purely moral considerations are therefore unknown, and it remains unclear whether these differ from those involved in the neural valuation of material goods.

Differences in these sets of processes are suggested by theoretical accounts emphasizing that moral preferences may originate from specific value-computation mechanisms. These accounts rest on the observation that many people perceive human lives as having an intrinsic (sacred) value (Sandel 2012; Dogan et al. 2016) that cannot, and should not, be measured on the same scale as the value of material objects (Kleinig 1991). For example, widespread outrage is usually observed when people realize that the value of human lives is explicitly quantified in terms of money, for instance during choices between health policies (Kmietowicz 2001), in the context of a company's decision on whether to re-call a dangerous car model (Dowie 1977), or when people are traded for money (Chuang 2006). Based on these observations, it has been proposed that assigning a financial value to a human life appears intuitively wrong for many people (Sandel 2012), suggesting that moral valuation may be implemented by processes that are distinct from those involved in the valuation of material goods.

In the present work, we test this alternative hypothesis by explicitly comparing the neural instantiation of moral and financial value-computations. We measured these with structurally equivalent choice tasks that differed only in the content of the choice-options: Human lives for moral decisions and monetary rewards for financial decisions. We decided to focus on human lives since subjective moral values are essential for the difficult decisions whether some lives are more valuable than others. One example are decisions about recipients of an organ transplant, for which it is

often required to implement a policy ranking among the potential recipients to decide who is most deserving to receive the organ (Courtney & Maxwell 2009). We adapted this decision situation to study the neural representations of subjective moral values, which we derived by fitting standard computational models of value-based decision making to the observed choices (Chung & Herrnstein 1967; Green et al. 2004; Rubinstein 2003) and correlating the estimated values with neural activity as measured with functional magnetic resonance imaging.

In order to fully capture individual behavioral variability during both decision types, we varied the decision-relevant characteristics of the choice options along two dimensions. For the financial decisions, participants chose between options that differed in terms of both the monetary amount and the temporal delay at which the amounts would be paid out. The subjective value of the choice options therefore depends inherently on individual time preferences (Green & Myerson 2004; Kable & Glimcher 2007; McClure 2004), which determine how the reward magnitude (i.e., the amount of money one can receive) is discounted by the delay (i.e., the number of days) one has to wait until receiving the reward. The moral decisions were constructed to match exactly this structure: They consisted of a customized moral scenario similar to the trolley moral dilemma (Foot 1967) that parametrically varied a choice-relevant magnitude (i.e., the number of lives one could save) that was discounted by a second factor which modulated the value of the lives at stake. This factor was the moral deservingness of the person that would have to be sacrificed in order to save the others (as indicated by different prior criminal records of this person). Both of these factors have been previously shown to play important roles in moral judgments (Shenhav & Greene 2010; Kliemann et al. 2008), but have never been combined in a choice setting as here.

We directly compared the neural value representations underlying both types of choices in the same participants using functional magnetic resonance imaging (fMRI). We ensured that the perceptual and sensorimotor demands required by both types of choices were kept similar, as the choice screens in both contexts were similarly arranged (Fig. 1a & b) and as responses were given with the same motor actions. Based on the existing value-based literature, we estimated subjective values underlying the financial choices by means of computational modeling (Frederick 2003; Rubinstein 2003) and expected to confirm their neural representations in brain activity in the vmPFC, the VS, and the PCC (Figner et al. 2010; Kable & Glimcher

2007; McClure 2004). We estimated moral subjective values with structurally isomorphic computational models; this allowed us to test whether moral subjective values would be represented by similar structures as financial values (e.g., the vmPFC, Shenhav & Greene 2010). Alternatively, moral values could engage specific representations (e.g., in the right TPJ, Kliemann et al. 2008; Young et al. 2007), thereby disproving the common currency hypothesis.

RESULTS

Behavioral Results

In both types of decisions, participants selected between two choice alternatives on each trial: For financial decisions (Fig. 1a), participants chose between 20 Swiss Francs (CHF) to be received today or an equal or larger financial reward (min = 20 CHF, max = 120 CHF) paid out after one of six different time delays (min = 1 day, max = 180 days). For moral decisions (Fig. 1b), participants chose between saving the lives of a larger number of people (min = 1, max = 10) at the expenses of sacrificing the life of one person, or not harming the one person and letting the group die. Moreover, closely mirroring the financial task, participants had to consider an associated feature that may discount the choice option´s value: the moral deservingness of the lives at stake, a property known to play an important role in modulating moral decisions (Kliemann et al. 2008). We implemented this by assigning one of six different prior criminal records (ranging from no criminal record to serial killer) to the single person that could be saved or harmed for the benefit of the group. Critically, our moral task did not require participants to read lengthy and complex moral vignettes before reporting decisions, but instead presented simple binary decisions where the only varying elements influencing the decisions were the two experimental variables magnitude and deservingness.

Subjective financial and moral values were estimated based on the participants' financial or moral choices respectively. In the reward domain, previous studies have repeatedly shown (Green & Myerson 2004; McClure 2004) that discount rates are typically well captured by hyperbolic functions, both in humans and other animals (Frederick et al. 2002; Green et al. 2004). In order to estimate participants' subjective financial values, we modeled the behavioral data with a standard hyperbolic function:

$$SV_f = LL / (1 + K_f * T) \quad (1)$$

Where SV_f is the subjective financial value of the delayed option estimated as fraction of the immediate reward, LL is the larger later amount offered, K_f corresponds to a subject-specific financial discounting constant, and *T* represents the time (in days) people had to wait to receive the reward. Consistent with previous findings (Kable & Glimcher, 2007), our participants' discounting curves were well modeled by this function (Fig. 1e; $R^2 = 0.98 \pm 0.015$). Moreover, the financial discount factors (K_f), and hence the SV_f, varied substantially across participants (ranging from $K_f = 3.78*10^{-5}$ to $K_f = 0.43$, fig. 1c).

Behavior in the moral task was modeled with a structurally equivalent model to the one used in the financial domain. This allows us to compare the estimated SV for each task both at the level of behavior (e.g., testing for a correlation between the two SV types) and brain activity. Specifically, behavior in the moral task was modeled with the following hyperbolic function:

$$SV_m = HL / (1 + K_m * D) \quad (2)$$

Where SV_m is the subjective moral value of saving the lives of the larger group by sacrificing the life of one person. The HL reflects the number of human lives one can save in the larger group; K_m corresponds to a subject-specific moral discount factor, and *D* represents the moral deservingness (i.e., criminal record) of the person one could sacrifice. As a first important result, we found that individual discount curves for moral choices (Fig. 1d) were well fit using equation 2 ($R^2 = 0.96 \pm 0.03$, Fig. 1f). This finding suggests that the moral subjective values estimated here indeed play an important role in moral decision making. Furthermore, like in the financial domain, moral subjective values and the moral discounting factors (K_m) varied substantially across participants (ranging from $K_m = 9.3*10^{-2}$ to $K_m = 7.08$).

While the two types of choices were comparable in terms of their computational requirements, they obviously differed qualitatively in terms of choice options and their consequences: On one hand, participants made decisions about whether or not to harm a human to save other lives, while on the other, they decided between different financial payoffs. It may therefore be expected that the two types of choices may differ in terms of response difficulty. However, the response times (RTs) for the two types of decisions – a standard proxy to measure task difficulty – were similar, as revealed by a t-test comparing RTs across the two tasks (average RTs moral 1214ms +/- 28 (s.e.m.), financial 1235ms +/- 24 (s.e.m.), $t(24) = 0.39$, $p = 0.7$). Moreover, we found no differences between the two tasks in how RTs varied as

a function of changes in the discounted values of the choice options (t(24) = 0.09, *p* = 0.92). Thus, the two types of decisions did not differ in terms of choice difficulty, which allows an unbiased comparison between the underlying neural mechanisms. Interestingly, although financial and moral choices were on average well fitted by identical functions and did not differ with respect to task-difficulty/RTs, we could not find behavioral evidence suggesting that moral and financial valuation processes rely on shared psychological mechanisms. When testing for a relationship between each individual's discounting in the financial and moral domain, we found no correlation between both discounting factors K_m and K_f (r = - 0.195, *p* = 0.35, Fig. 1g). This absence of a correlation already seems to suggest that moral and financial value estimations may be performed by independent neural decision mechanisms.

Functional Imaging Results

As an initial imaging analysis step, we confirmed the well-known neural correlates of subjective financial values. As expected based on the literature (Levy & Glimcher 2011; Bartra et al. 2013; Kable & Glimcher 2007), we found a significant correlation between subjective financial values of the delayed monetary option with blood oxygen level-dependent (BOLD) activity in brain areas associated with subjective financial value-processing (Grueschow et al. 2015; Clithero & Rangel 2014). In particular, we found the hypothesized financial subjective value representations in the vmPFC, dmPFC, and PCC (Fig. 2a), as well as in the VS (small-volume-corrected P < 0.05, see Fig. 2a and Table 1).

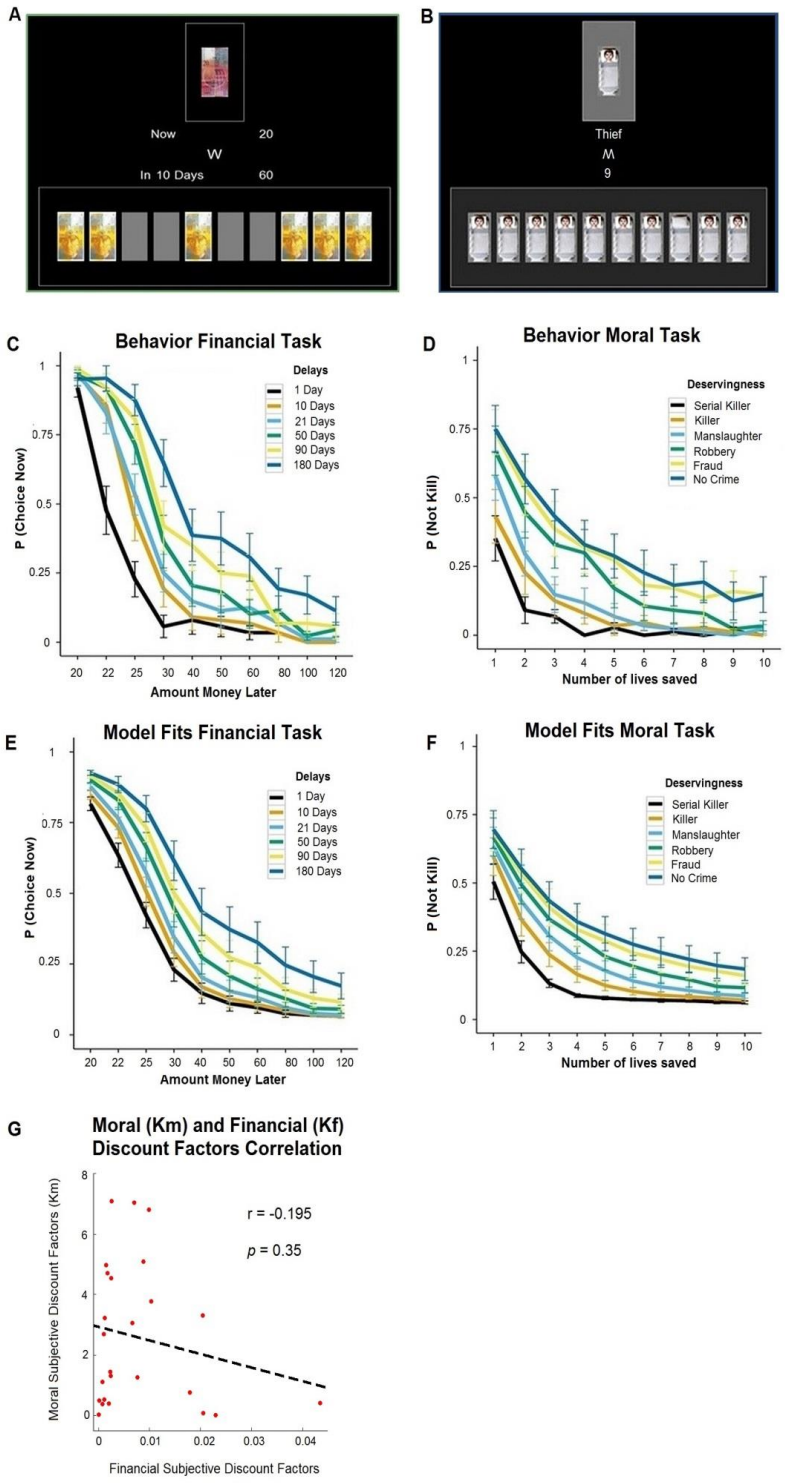


Figure 1: Paradigm and Behavioral Results: Participants made financial (**a**) and moral (**b**) choices. In the financial task, they decided whether or not to give up a sooner smaller financial reward for a later larger financial reward. Similarly, in the moral task, they decided whether or not to sacrifice one coma-patient to save a larger group of accident victims requiring organ transplants. (**c**) The probability of giving up the sooner smaller reward increased as the amount of the delayed reward increased. The increase was modulated by the delay participants had to wait to receive the larger option. (**d**) Similarly, the probability of killing the one person in order to save the larger group of people increased with the number of people that could be saved; in this case the probability of choosing to sacrifice the coma-patient was modulated by deservingness. Behavior in both tasks was well captured by the models used, as revealed by the model fits for the financial **(e)** and the moral **(f)** task. Although choice in both types of decision tasks was well modelled by structurally equivalent models, we found no evidence of correlation between financial and moral discounting (**g**).

More importantly, our fMRI analysis also revealed a set of
comprising the bilateral TPJ, the PCC, the right DLPFC, the ante
inferior parietal lobule (IPL), and the anterior cingulate cortex (ACC
signal represented subjective moral values (see figure 2b and 2c, a

TABLE 1

Region	Peak-Side	Cluster Size	x	y	z	Z score	T score	p-value
Neural Correlates of Subjective Moral Values								
ACC		839	0	29	40	5.14	7.06	<0.001
AntIns	R	164	36	2	10	3.78	4.48	<0.001
Cuneus	R	150	15	-88	4	4	4.83	<0.001
DLPFC	R	839	48	23	43	4.86	6.44	<0.001
IPL	L	105	-51	-37	25	4.62	5.94	<0.001
IPL	R	110	60	-16	22	4.28	5.32	<0.001
PCC		489	0	-67	37	5.22	7.23	<0.001
TPJ	R	518	48	-58	31	4.59	5.89	<0.001
Neural Correlates of Subjective Financial Values								
DMPFC	L	1831	-9	50	43	6.39	10.57	<0.001
MTG	L	238	-57	-7	-14	4.44	5.6	<0.001
PCC		194	0	-49	31	4.69	6.08	<0.001
SMG	R	273	63	-25	1	4.34	5.42	<0.001
STS	L	597	-57	-37	25	4.53	5.74	<0.001
STS	R	139	45	-28	22	3.64	4.28	<0.001
Visual Cortex	R	1777	12	-85	4	5.87	8.92	<0.001

Table 1: Average brain activity explicitly representing subjective moral values (rows 4-11), related to Figure 2B and 2C, and average brain activity explicitly representing subjective financial values (rows 13-19), related to Figure 2A.
All p-values are FWE-corrected for the whole brain. ACC = anterior cingulate cortex; AntIns = Anterior Insula; DLPFC = dorsolateral prefrontal cortex; IPL = Inferior parietal lobule; PCC = posterior cingulate cortex; TPJ = temporo-parietal junction; DMPFC = dorsomedial-prefrontal cortex; MTG = medial temporal gyrus ; SMG = supramarginal gyrus ;STS = superior temporal

to differences in neural activity in these brain regions, effectively providing novel evidence of a neural signature of subjective moral preferences. More specifically, we found that the higher the subjective moral value of a human life, the higher the BOLD activity in the rTPJ, the PCC and the right DLPFC (Fig. 2c). Moreover, our results show that a decrease in the subjective value of a human life (and therefore an increase in the tendency to sacrifice this person to save a larger group) was

represented by increasing neural activity in the bilateral anterior insula, the left inferior parietal lobule (IPL), and the anterior cingulate cortex (ACC). These results are generally consistent with previous reports of activity in these brain areas during moral decisions (C. A. Hutcherson et al. 2015; Kliemann et al. 2008; Greene et al. 2004; Greene et al. 2001), as well as in the representation of expected values in the moral domain (Shenhav & Greene 2010).

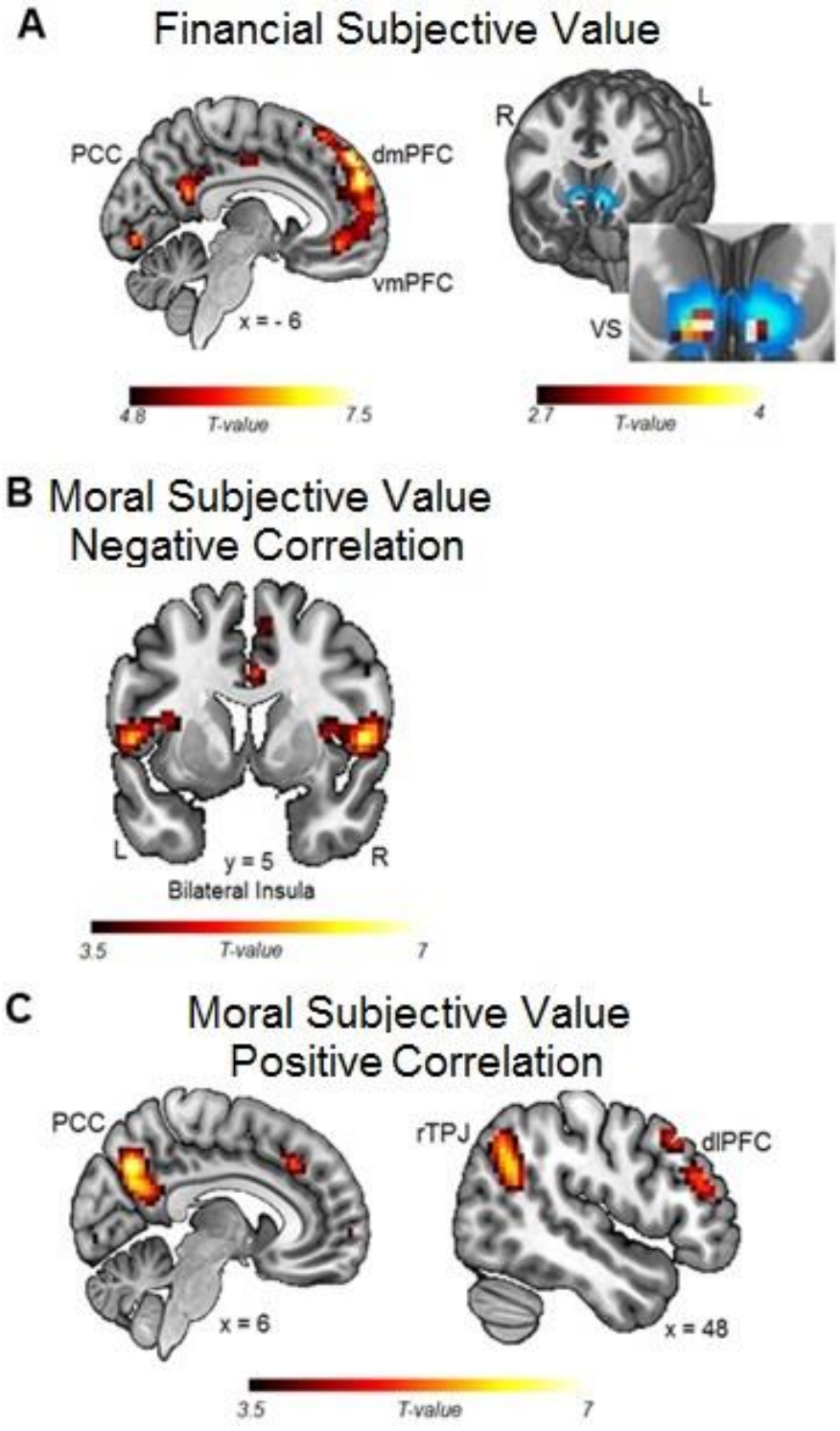


Figure 2

Functional Imaging Results. (**a**) Financial subjective values were represented by neural activity in the mPFC, the PCC, and the VS, consistent with previous findings; Cyan area (right) corresponds to the nucleus accumbens volume mask provided by the FSL-Harvard-Oxford-atlas. (**b, c**) Moral subjective values were represented by neural activity in the bilateral

Another aim of our fMRI analysis was to establish if moral subjective value computations rely on domain-general mechanisms also shared with non-moral value-based decisions (Shenhav & Greene 2010) or whether they rely on markedly different brain regions. When directly comparing the neural activity related to moral subjective value computations versus that related to the matched financial value computations, we found that moral subjective values were more strongly related to activity in the rTPJ and the PCC (Fig. 3 and Table 2). In contrast, neural activity in the dmPFC (Fig. 3 and Table 2) was more strongly involved in representing financial than moral values. These findings highlight that the neural representation of subjective moral values relies largely on domain-specific mechanisms. We also tested for potential functional activity involved in computing both moral and financial subjective values. The conjunction analysis testing for such overlap, however, did not reveal any significant result.

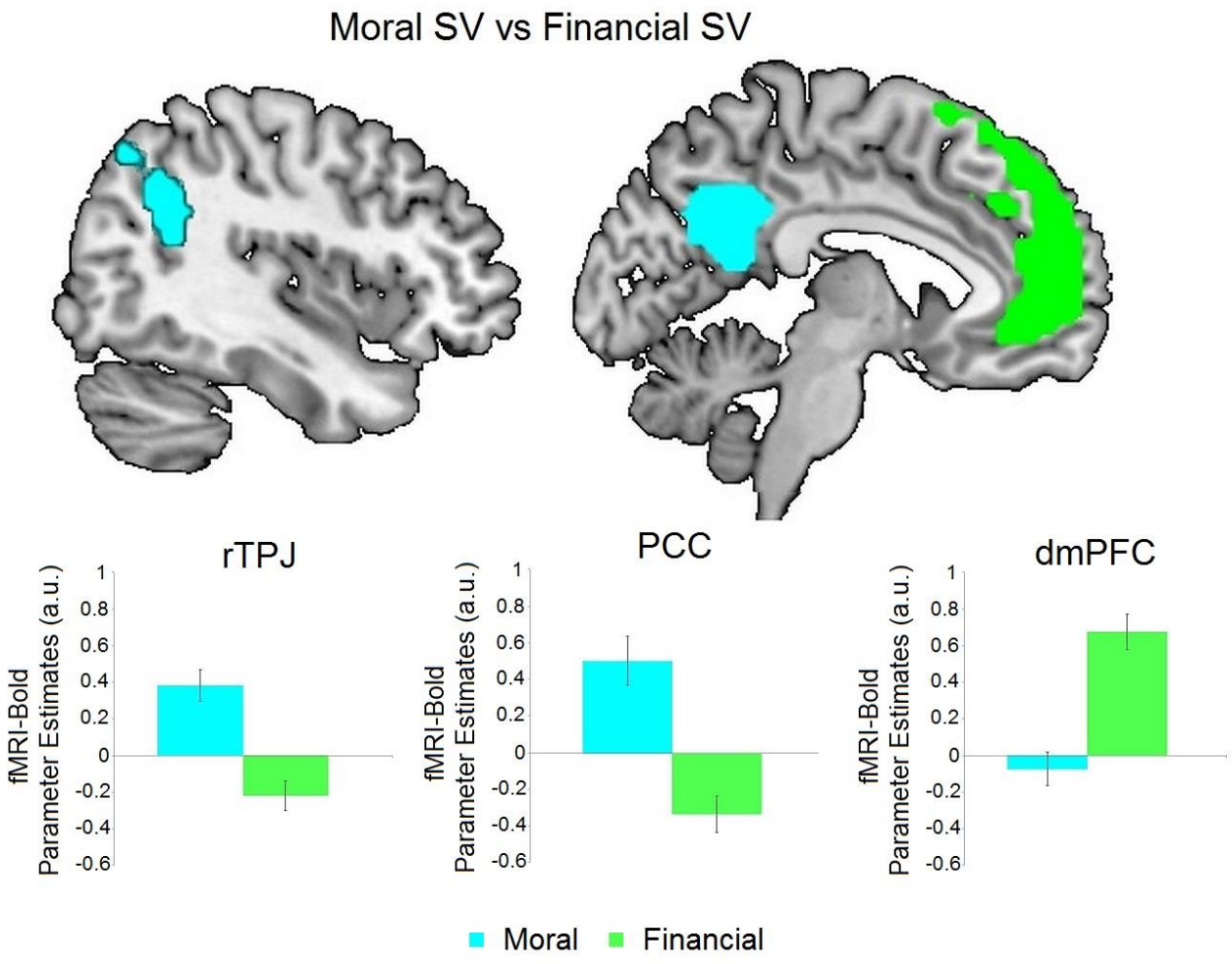


Figure 3

Domain specific subjective value representations: We found neural representations specific to moral subjective value in the rTPJ and the PCC (cyan). In contrast, financial-specific subjective value correlates were instead identified in the dmPFC (green).

TABLE 2								
Region	Peak-Side	Cluster Size	x	y	z	Z score	T score	p-value
Neural Correlates of Subjective Moral Values > Subjective Financial Values								
Cuneus	R	1908	15	-88	4	6.35	10.44	<0.001
PCC		471	0	-55	25	5.25	7.31	<0.001
TPJ	R	283	57	-64	28	5.02	6.78	<0.001
Neural Correlates of Subjective Financial Values > Subjective Moral Values								
DMPFC	L	1303	-15	47	40	5.56	8.07	<0.001
STS	L	347	-36	-61	25	5.16	7.09	<0.001

Table 2: Average brain activity specifically representing subjective moral values > subjective financial values (rows 4-6), and average brain activity specifically representing subjective financial values > subjective moral values (rows 8-9), related to Figure 3.
All p-values are FWE-corrected for the whole brain. ACC = anterior cingulate cortex; PCC = posterior cingulate cortex; TPJ = temporo-parietal junction; DMPFC = dorsomedial-prefrontal

The main motivation for this study was to identify neural value representations that underlie individual moral preferences, thereby testing if the brain represents moral and material preferences on a common neural currency. This hypothesis is consistent with widely held views on the domain-generality of neural value processes, but contradicts the moral intuition that human lives should not be valued in material terms and on the same currency as objects. We tested the hypothesis with a novel moral choice paradigm allowing us to a) investigate if the human brain explicitly represents estimations of the subjective value of saving/harming the life of other persons and b) testing whether these neural representations differ from those representing financial subjective value. Our behavioral models show that, similar to financial choices, moral decisions concerning who should be saved/harmed are well fit by computational decision models estimating the subjective value that people assign to the choice alternatives. Neurally, we not only found that moral subjective values are computed according to similar principles as financial values, but more importantly we have provided evidence that these computations are instantiated in domain-specific brain areas.

Our data shows that subjective moral values are explicitly represented in a network of regions comprising the right TPJ, the PCC, the right DLPFC, the left IPL, and in

the anterior insula. More importantly, directly comparing the neural activity elicited by moral vs. financial subjective values allowed us to demonstrate that the moral subjective values of human lives are not represented by the same common neural currency representing the value of material types of goods (Levy & Glimcher 2012; Izuma et al. 2008; Zink et al. 2008). Instead, we found that neural activity in the rTPJ, PCC, and other areas (see Figure 3 and Table 2), was specifically involved in the representation of moral compared to financial subjective values, suggesting the existence of a moral-specific valuation system eliciting neural activity in brain areas previously implicated in representing morally salient components of decisions, such as empathy, harm-aversion or, as shown in the present paper, estimations of lives saved and harmed. Thus, our findings are at odds with the assumption that moral decisions rely on domain-general decision processes (Shenhav & Greene 2010). This latter view is mainly supported by an indirect comparison of the neural activity elicited by moral and financial expected value computations. In their study, the authors (Shenhav & Greene 2010) found that computations of the expected value of probabilistic outcomes in moral scenarios elicited neural activity in regions commonly associated with computations of the expected value of probabilistic financial rewards, such as the striatum (Tobler et al. 2006) and the vmPFC (Hare et al. 2008; Knutson & Peterson 2005). However, given the absence of a direct comparison with a monetary control task, it was impossible to disentangle if the identified neural activity in (Shenhav & Greene 2010) is associated only with calculations of the expected value of outcomes – irrespective of the context – or if it is also processing morally relevant information.

Furthermore, studies investigating decisions in contexts that require integrating moral and financial values, (e.g., deciding between donating to a charity or keeping the money for oneself) found that the subjective value of choice options was mostly represented in the vmPFC but modulated by social information from the TPJ (Hare et al. 2010; Soutschek et al. 2016; Strombach et al. 2015; Crockett et al. 2014). In this decision context, the rTPJ has been thought to estimate socially salient components, such as the need to overcome one's perspective and the deservingness of a charity, and to pass this information on to the vmPFC, where the value-computation is ultimately implemented. Our results not only complement the existing literature isolating the neural representations dedicated specifically to computing moral subjective values, but offer also a novel and intriguing perspective on the role of the

rTPJ in moral value-computations. For decisions based only on moral values (i.e., where there is no trade-off between self-interested financial values and moral values), our evidence suggests that subjective moral values can be represented directly in the rTPJ, without any vmPFC involvement.

Relating different moral preferences to neural activity, we found that the subjective value of a human life is associated with neural activity in the rTPJ, PCC, and DLPFC among other areas (see Table 1 and Figure 1c). In contrast, we found a set of regions, including the Anterior Insula and the left IPL, that displayed a negative relation to the estimated subjective value of a human life (see Table 1 and Figure 1c). These findings suggest an intriguing novel mechanistic interpretation of how moral preferences are represented in the brain. It is plausible that the subjective value of a human life relies on processes that measure the harm inflicted to others, consistently with previous studies linking brain activity in the right TPJ, PCC, and DLPFC to processing harm aversion and empathy (Ugazio et al. 2014; Majdandžić et al. 2012; Crockett et al. 2010; Crockett et al. 2017). Conversely, the moral preference that considers required saving a larger number of people relies on neural valuation mechanisms responsible for comparing the magnitudes of the moral choice options, reflected in brain activity in the left IPL and in the anterior insula. This interpretation accommodates and extends the ideas proposed in a previous study that identified a positive correlation between these bran areas and an increase of expected value in moral decisions with probabilistic outcomes (Shenhav & Greene 2010).

The analysis of the monetary control task showed that financial subjective values were indeed represented by neural activity in the vmPFC, PCC, and VS, consistent with numerous previous findings (Kable & Glimcher 2007). Domain-general value-computation mechanisms may contribute to moral decisions (C. A. Hutcherson et al. 2015; Shenhav & Greene 2010), at least to the extent that financial valuation mechanisms can corrupt human moral values (Falk & Szech 2013) or that moral values related to the aversion of harming others can discount financial values (Crockett et al. 2017). However, our data critically revealed a dissociation in the neural networks involved in the representations of purely moral subjective values and those involved in the representation of financial subjective values. This raises the interesting question for future studies what context factors may determine whether or not domain-general valuation mechanisms are involved in moral choices.

More generally, while trolley-type moral dilemmas have been questioned for their ecological validity (Kahane 2015), recent technological developments in robotics and artificial intelligence have revitalized the importance of this type of dilemmas (Bonnefon et al. 2016). Our results may prove critical for informing future ethical, public and scientific debates regarding these technologies. For instance, it is increasingly debated how a self-driving car should be pre-programmed for selecting whom to harm in potentially critical situations where different lives are at stake - should it always protect the people inside the car or should it use some other criterion? Our current results identify distinct neural mechanisms by which our brains compute tradeoffs between saving and harming human lives, which differ from neural valuation processes involved in selecting between material goods. This suggests that artificial intelligence would need to account for the properties of these mechanisms in order to be perceived as morally appropriate. Last but not least, our study illustrates how moral preferences may be assessed in a manner that is computationally similar to the assessment of financial preferences, without requiring the participants to read and understand complex moral vignettes. This facilitates identification of the choice-related brain mechanisms and may prove essential for a move towards an integrated perspective of how the brain controls and integrates moral and material concerns in the control of actions.

EXPERIMENTAL PROCEDURES

Participants

The participants were twenty-five healthy students from the University of Zurich (age: min 19, max 34, mean = 22.08, S.E.M. = 0.74 years old; 13 females) with no reported history of neurological or psychiatric disorder and no current use of medication as measured with standard surveys. All the experimental procedures were approved by the Research Ethics Committee of the Canton of Zurich.

fMRI Task

Participants made financial and moral choices in randomly alternating blocks during the event-related fMRI sessions. Visual stimulation was highly similar (**cf. Fig 1a and 1b**) while the required motor commands where identical. Both tasks were cued visually (a ‘W’ cued financial trials, while moral trial were cued by the letter ‘M’).

During financial choice trials participants indicated whether they preferred to give up a reward of 20 CHF that was paid out immediately after the study in order to receive a variable reward (20, 22, 25, 30, 40, 50, 60, 80, 100 or 120 CHF) after waiting

different amounts of days (1, 10, 21, 50, 90, or 180 days). One financial trial was randomly selected at the end of the study and paid out to the participant as described above. If a delayed option was selected, the money was sent via mail to the address specified by the participant.

The moral task required participants to read a moral scenario before starting the fMRI session. This moral scenario instructed them to place themselves in the shoes of a doctor who is taking care of a different patient in a coma-state every day. The moral deservingness of these patients differed as indicated by different criminal records (no records of criminal activities, fraud, robbery, manslaughter, killing a person, killing multiple persons). While on duty, this doctor is informed about the sudden need of organ transplants in variable amounts of victims due to an accident (from a minimum of 1 to a maximum of 10 victims). These people would die if they did not receive the organs soon.

The participant is then asked what she/he is morally required to do in the shoes of the doctor: choice alternative 1) interrupt life-support to the coma patient, resulting in his death, and use the organs to save the lives of the accident victims; choice alternative 2), leave the coma-patient on life-support and let the victims of the accident die. During the moral choice trials, participants reported which course of action was the morally required one. Each trial consisted of a unique combination of deservingness and number of people the doctor could save by harming the coma-patient.

Behavioral analysis

Reaction times were analyzed with a two-sided paired t-test comparing the individual average RTs in the moral compared to the financial task. The relationship of RTs with choice was analyzed with a two-sided t-test comparing the standardized slopes (β_1) of a linear regression (formally, RTs = β_0 + β_1SV + E) estimating the relation between RTs and moral and financial subjective values for each individual.

Financial and moral subjective values were estimated using structurally identical models. Respectively, financial subjective values were estimated with the model:

$$P_{\text{choice}} = \frac{1}{1+\exp(-(B_0+B_1\times[20-\text{SVf}(kf)]} \quad (3)$$

where the function $SV_f(k_f)$ is the subjective value for the financial choice (defined in Eq. 1, see Results section) and the parameter k_f corresponds to the discount factor of the hyperbolic function; Moral subjective values were estimated with the model:

$$P_{\text{choice}} = \frac{1}{1+\exp(-(B_0+B_1\times[1-\text{SVm}(km)]} \quad (4)$$

where the function $SV_m(k_m)$ is the subjective value for the moral choice (defined in Eq. 2, see Results section) and the parameter k_m corresponds to the discount factor of the hyperbolic function.

In both cases, the fitting strategy was based on a Bayesian Hierarchical Modeling (BHM) approach. This approach constitutes an attractive compromise between the extremes of complete pooling and complete independence (Gelman & Hill 2007). As in the complete independence approach, BHM estimates parameters for each individual participant. However, these estimates avoid the averaging artifacts that come with the complete pooling approach as well as the unreliability that comes with the estimation of parameters for individual participants. A Bayesian model was fit for each choice type (moral and financial) and contained random effects for the three subject-specific parameters of interest and assumed flat priors for all parameters at the highest hierarchy level. Inference of the parameters in the BHM was performed via the Gibbs sampler using the Markov Chain Montecarlo (MCMC) technique implemented in JAGS (Polanía et al. 2015; Plummer 2003). A total of 10,000 samples were drawn from an initial burn-in step and subsequently a total of 10,000 new samples were drawn with three chains (each chain was derived based on a different random number generator engine, and each with a different seed). We applied a thinning of 10 to this final sample, thus resulting in a final set of 1,000 samples for each parameter. This thinning assured that the final samples were auto-decorrelated for all of the latent variables of interest. We conducted Gelman–Rubin tests for each parameter to confirm convergence of the chains. All latent variables in our Bayesian models had $\widehat{\mathrm{R}} < 1.05$, which suggests that all three chains converged to a target posterior distribution.

fMRI data-acquisition and pre-processing.

Subjects performed four choice-task-sessions (each containing 60 financial and moral perceptual choices) and one resting-state-session that lasted 6.5 minutes each. During each session, we acquired 270 T2*-weighted whole-brain echo planar images using a Philips Achieva 3 T whole-body scanner (Philips Medical Systems, Best, The Netherlands) equipped with an 8-channel Philips sensitivity-encoded (SENSE) head coil. Imaging parameters were: 2600 ms repetition time (TR); 37 slices (transversal, ascending acquisition); 2.6 mm slice thickness; 2.5 mm x 2.5 mm

in-plane resolution; 0.65 mm gap; 90° flip angle. To measure at fully equilibrated magnetic field, five dummy image excitations were performed and discarded before functional image acquisition started. To enhance BOLD-contrast sensitivity throughout the brain, we used a dual-echo-sequence (TE: 17 ms and 44 ms) in combination with a weighted voxel-wise summation technique (Posse et al., 1999; Schmiedeskamp et al., 2010) that generates a single functional whole-brain image with optimal sensitivity for each TR. For this procedure, the signal-to-noise ratio is first computed for each echo image voxel in the resting-state scan. These SNR measures are then used to weight each voxel in the two echo images acquired per TR of the choice-task sessions according to the formula

$$X = \frac{X_{E1} \cdot SNR_{E1} + X_{E2} \cdot SNR_{E2}}{SNR_{E1} + SNR_{E2}}$$

where X is the resulting image for a given TR, X_{E1} and X_{E2} are the images acquired at that TR for the first echo and second echo, respectively, and SNR_{E1} and SNR_{E2} are the signal-to-noise images (generated as voxel-wise mean divided by the voxel-wise standard deviation) for the resting-state time-series acquired for the first echo and second echo, respectively. A high-resolution T1-weighted whole brain structural image used for image registration during post-processing (181 sagittal slices; matrix size: 256 x 256; voxel size: 1 x 1 x 1 mm ; TR/TE/TI: 8.3/2.26/181 ms) was also acquired for each subject.

Image preprocessing and analysis were conducted using SPM8 (Wellcome Trust Centre for Neuroimaging). Functional images were slice-time corrected (to the middle slice acquisition time) and realigned (accounting for individual head motion). Each participant's T1-weighted structural image was co-registered with the mean functional image and normalized to the standard T1 MNI template using the new-segment-procedure provided by SPM8 (Ashburner & Friston 2005). The functional images were then normalized to the standard MNI template using the same transformation, spatially resampled to 3 mm isotropic voxels, and smoothed using a Gaussian kernel (FWHM, 8mm).

fMRI data-analysis

The general linear model (GLM) we implemented was suited to identify and contrast correlations of BOLD signals with financial and moral subjective values during the financial and the moral trials respectively. The included regressors were therefore (1) an indicator function for financial choices with (2) financial subjective value (z-scored

at the individual level) as parametric modulator, (3) an indicator function for moral choices with (4) the parametric modulator moral subjective value (z-scored at the individual level). Two additional orthogonalized parametric modulators were also included: a first one for the objective delays (financial task) and a second one for the deservingness levels (moral task). Our regressors of interest were modelled as stick-functions at the time of stimulus onset convolved with a canonical hemodynamic response function; these regressors were regressed against the blood oxygen level-dependent (BOLD) signal in each voxel. In addition to these main regressors, the GLM also included several regressors of no interest: two indicator functions for financial and moral block cues, and six motion parameters (obtained during the realignment procedure).

First-level summary statistics were obtained by calculating single-subject voxel-wise contrasts for each of the two subjective value parametric modulators. Second-level random effects group contrast maps were then tested for significance by one-sample t tests across single-subject contrast maps. Statistical inference was performed at the cluster level, using a whole-brain FWE-corrected statistical threshold of $P < 0.05$ (based on a cluster-forming voxel cut-off set to $P < 0.001$). For the hypothesis-guided ROI analysis of the ventral striatum (VS), we corrected for multiple comparisons using a small-volume correction (SVC, $P < 0.05$) within the bilateral nucleus accumbens volume mask provided by the FSL-Harvard-Oxford atlas.

AUTHOR CONTRIBUTIONS

G.U. and C.C.R. conceived the study.

G.U., M.G., C.L., P.N.T., and C.C.R. designed the study.

G. U., performed experiments

G. U., M.G., and R.P. analyzed data with conceptual input from C.L., P.N.T., and C.C.R.

G. U., M.G., R.P., C.L., P.N.T., and C.C.R. wrote the manuscript.

ACKNOWLEDGEMENTS

We thank Karl Treiber for scanning assistance. This work was supported by grants of the Swiss National Science Foundation (PBZHP1_147240, PP00P1_128574, PP00P1_150739, 00014_165884, 105314_152891, 320030_143443, and CRSII3_141965**)** to G.U., P.N.T, and C.C.R. All authors gratefully acknowledge support by the Zurich Center for Neurosciences (ZNZ).

References

Ashburner, J. & Friston, K.J., 2005. Unified segmentation. *NeuroImage*, 26(3), pp.839–851.

Bartra, O., McGuire, J.T. & Kable, J.W., 2013. The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, 76, pp.412–427. Available at: http://dx.doi.org/10.1016/j.neuroimage.2013.02.063.

Berns, G.S. & Atran, S., 2012. The biology of cultural conflict. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1589), pp.633–639.

Bonnefon, J.-F., Shariff, A. & Rahwan, I., 2016. The social dilemma of autonomous vehicles. *Science*, 352(6293), pp.1573–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27339987.

Chuang, J., 2006. Beyond a Snapshot: Preventing Human Trafficking in the Global Economy. *Indiana Journal of Global Legal Studies*, 13(1), pp.137–163. Available at: http://muse.jhu.edu/journals/gls/summary/v013/13.1chuang.html.

Chung, S.H. & Herrnstein, R.J., 1967. Choice and delay of reinforcement. *Journal of the experimental analysis of behavior*, 10(1), pp.67–74. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1338319&tool=pmcentrez&rendertype=abstract.

Clithero, J.A. & Rangel, A., 2014. Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, 9(9), pp.1289–1302. Available at: http://scan.oxfordjournals.org/lookup/doi/10.1093/scan/nst106.

Courtney, A.E. & Maxwell, A.P., 2009. The challenge of doing what is right in renal transplantation: Balancing equity and utility. *Nephron - Clinical Practice*, 111(1).

Crockett, M.J. et al., 2014. Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences of the United States of America*, 111(48), pp.17320–5. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4260587&tool=pmcentrez&rendertype=abstract.

Crockett, M.J. et al., 2017. Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, (May), pp.1–10. Available at: http://dx.doi.org/10.1038/nn.4557.

Crockett, M.J. et al., 2010. Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences*, 107(40), pp.17433–17438. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1009396107.

Dogan, A. et al., 2016. Prefrontal connections express individual differences in intrinsic resistance to trading off honesty values against economic benefits. *Scientific Reports*, 6.

Dowie, M., 1977. Pinto Madness. *Mother Jones*, (September/October), pp.1–15. Available at: http://www.motherjones.com/politics/1977/09/pinto-madness.

Falk, A. & Szech, N., 2013. Morals and Markets. *Science*, 340(6133), pp.707–711. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1231566.

Figner, B. et al., 2010. Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience*, 13(5), pp.538–539. Available at: http://www.nature.com/doifinder/10.1038/nn.2516.

Foot, P., 1967. The Problem of Abortion and the Doctrine of the Double Effect. *OxFoot, P., 1967. The Problem of Abortion and the Doctrine of the Double Effect. Oxford Review, (5), pp.5–15.ford Review*, (5), pp.5–15.

Frederick, S., 2003. Measuring Intergenerational Time Preference: Are Future Lives Valued Less? *Journal of Risk and Uncertainty*, 26(1), pp.39–53.

Frederick, S., Loewenstein, G. & O'Donoghue, T., 2002. Time Discounting and Time PreferenceL A Critical Review. *Journal of Economic Literature*, 40, pp.351–401. Available at: http://www.jstor.org/stable/10.2307/2698382%5Cnpapers3://publication/uuid/4275A130-0862-4B75-9474-5DF36AE7420A.

Gelman, A. & Hill, J., 2007. *Data analysis using regression and multilevel/hierarchical models*,

Green, L. et al., 2004. Discounting of delayed food rewards in pigeons and rats: is there a magnitude effect? *Journal of the experimental analysis of behavior*, 81(1), pp.39–50. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1284970&tool=pmcentrez&rendertype=abstract.

Green, L. & Myerson, J., 2004. A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, 130(5), pp.769–792.

Greene, J.D. et al., 2004. The Neural Bases of Cognitive Conflict and Control in Moral Judgment. , 44, pp.389–400.

Greene, J.D., Sommerville, R.B. & Nystrom, L.E., 2001. An fMRI Investigation of

Emotional Engagement in Moral Judgment. , 293(September), pp.2105–2108.

Grueschow, M. et al., 2015. Automatic versus Choice-Dependent Value Representations in the Human Brain. *Neuron*, 85(4), pp.874–885.

Hare, T. a et al., 2008. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(22), pp.5623–5630.

Hare, T. a et al., 2010. Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(2), pp.583–590.

Hutcherson, C.A. et al., 2015. Emotional and Utilitarian Appraisals of Moral Dilemmas Are Encoded in Separate Areas and Integrated in Ventromedial Prefrontal Cortex. *Journal of Neuroscience*, 35(36), pp.12593–12605. Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3402-14.2015.

Hutcherson, C.A. et al., 2015. Emotional and Utilitarian Appraisals of Moral Dilemmas Are Encoded in Separate Areas and Integrated in Ventromedial Prefrontal Cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 35(36), pp.12593–605. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26354924%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4563040.

Izuma, K., Saito, D.N. & Sadato, N., 2008. Processing of Social and Monetary Rewards in the Human Striatum. *Neuron*, 58(2), pp.284–294.

Kable, J.W. & Glimcher, P.W., 2007. The neural correlates of subjective value during intertemporal choice. *Nature neuroscience*, 10(12), pp.1625–33. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2845395&tool=pmcentrez&rendertype=abstract.

Kahane, G., 2015. Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, (July), pp.1–10. Available at: http://www.tandfonline.com/doi/abs/10.1080/17470919.2015.1023400.

Kleinig, J., 1991. *Valuing life*,

Kliemann, D. et al., 2008. The influence of prior record on moral judgment. *Neuropsychologia*, 46(12), pp.2949–2957. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18606175.

Kmietowicz, Z., 2001. Tobacco company claims that smokers help the economy. *BMJ*, 323(7305), p.126. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11463670.

Knutson, B. & Peterson, R., 2005. Neurally reconstructing expected utility. *Games and Economic Behavior*, 52(2), pp.305–315.

Lamm, C. et al., 2007. What are you feeling? Using functional magnetic resonance imaging to assess the modulation of sensory and affective responses during empathy for pain. *PLoS ONE*, 2(12).

Levy, D.J. & Glimcher, P.W., 2011. Comparing Apples and Oranges: Using Reward-Specific and Reward-General Subjective Value Representation in the Brain. *The Journal of Neuroscience*, 31(41), pp.14693–14707. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3763520/%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC3763520/pdf/zns14693.pdf.

Levy, D.J. & Glimcher, P.W., 2012. The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, 22(6), pp.1027–1038.

Majdandžić, J. et al., 2012. The Human Factor: Behavioral and Neural Correlates of Humanized Perception in Moral Decision Making. *PLoS ONE*, 7(10).

McClure, S.M., 2004. Separate Neural Systems Value Immediate and Delayed Monetary Rewards. *Science*, 306(5695), pp.503–507. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1100907.

Plummer, M., 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, pp.20–22.

Polanía, R. et al., 2015. The precision of value-based choices depends causally on fronto-parietal phase coupling. *Nature Communications*, 6, p.8090. Available at: http://www.nature.com/doifinder/10.1038/ncomms9090.

Rubinstein, A., 2003. "Economics and psychology"? The case of hyperbolic discounting. *International Economic Review*, 44(4), pp.1207–1216.

Sandel, M.J., 2012. What Isn't for Sale? *The Atlantic*. Available at: http://www.theatlantic.com/magazine/archive/2012/04/what-isnt-for-sale/308902/.

Schultz, W., 2006. Behavioral theories and the neurophysiology of reward. *Annual review of psychology*, 57, pp.87–115. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16318590.

Shenhav, A. & Greene, J.D., 2010. Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude. *Neuron*, 67(4), pp.667–677. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0896627310005817.

Silani, G. et al., 2013. Right Supramarginal Gyrus Is Crucial to Overcome Emotional Egocentricity Bias in Social Judgments. *Journal of Neuroscience*, 33(39),

pp.15466–15476. Available at: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1488-13.2013.

Soutschek, A. et al., 2016. Brain stimulation reveals crucial role of overcoming self-centeredness in self-control. *Science Advances*, 2(October 19), pp.2–10.

Strombach, T. et al., 2015. Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences USA*, 112(5), pp.1619–1624. Available at: http://www.pnas.org/content/112/5/1619%5Cnhttp://www.pnas.org/lookup/doi/10.1073/pnas.1414715112.

Tobler, P.N. et al., 2006. Reward Value Coding Distinct From Risk Attitude-Related Uncertainty Coding in Human Reward Systems. *Journal of Neurophysiology*, 97(2), pp.1621–1632. Available at: http://jn.physiology.org/cgi/doi/10.1152/jn.00745.2006.

Ugazio, G., Majdandžić, J. & Lamm, C., 2014. Are Empathy and Morality Linked? Insights from Moral Psychology, Social and Decision Neuroscience, and Philosophy. *Empathy in Morality*, pp.155–171.

Young, L. et al., 2007. The neural basis of the interaction between theory of mind and moral judgment. , 104(20).

Zink, C.F. et al., 2008. Know Your Place: Neural Processing of Social Hierarchy in Humans. *Neuron*, 58(2), pp.273–283.

B. Appendix to Study 2

Increasing Honesty in Humans with Electrical Brain Stimulation

Alain Cohn[1*], Michel André Maréchal[2*], Giuseppe Ugazio[3], and Christian C. Ruff[2*]

A version of this paper is published as Marechal, M., Cohn, A., Ugazio, G., and Ruff, C. C. Increasing Honesty in Humans with Electrical Brain Stimulation. *PNAS*. 2017 114 (17) 4360-4364, 2017

Affiliations:

[1]Booth School of Business, University of Chicago, Chicago, IL 60637, Unites States

[2]Department of Economics, University of Zurich, 8006 Zurich, Switzerland

[3]Department of Psychology, Harvard University, Cambridge, MA 02138, Unites States

[*]Correspondence to: alain.cohn@chicagobooth.edu, michel.marechal@econ.uzh.ch, or christian.ruff@econ.uzh.ch.

Honesty plays a key role in social and economic interactions and is crucial for societal functioning. However, breaches of honesty are pervasive and cause significant societal and economic problems that can affect entire nations. Despite its importance, remarkably little is known about the neurobiological mechanisms supporting honest behavior. We demonstrate that honesty can be increased in humans with transcranial direct current stimulation (tDCS) over the right dorsolateral prefrontal cortex (rDLPFC). Participants (N=145) completed a die-rolling task where they could misreport their outcomes to increase their earnings, thereby pitting honest behavior against personal financial gain. Cheating was substantial in a control condition, but decreased dramatically when neural excitability was enhanced with tDCS. This increase in honesty could not be explained by changes in material self-interest or moral beliefs and was dissociated from participants' financial risk-taking, impulsivity, and mood. A follow-up experiment (N=156) showed that tDCS only reduced cheating when dishonest behavior benefited the participants themselves rather than another person, suggesting

that the stimulated neural process specifically resolves conflicts between honesty and material self-interest. Our results demonstrate that honesty can be strengthened by non-invasive interventions and concur with theories proposing that the human brain has evolved mechanisms dedicated to control complex social behaviors.

Keywords: honesty | cheating | experiment | transcranial direct current stimulation | dorsolateral prefrontal cortex

Significance:

Honesty affects almost every aspect of social and economic life. We conducted experiments in which participants could earn considerable amounts of money by cheating on a die-rolling task. Cheating was substantial but decreased by more than half during transcranial direct current stimulation (tDCS) over the right dorsolateral prefrontal cortex (rDLPFC). This stimulation-induced increase in honesty was functionally specific: It did not affect other types of behavioral control related to self-interest, risk-taking and impulsivity. Moreover, cheating was only reduced when it benefitted the participants themselves rather than another person. Thus, the human brain implements specialized processes that enable us to remain honest when faced with opportunities to cheat for personal material gain. Importantly, these processes can be strengthened by external interventions.

Main Text:

Dishonest behavior is pervasive and carries important economic and societal consequences (1–6). For example, illegal tax evasion is thought to account for over 5% of the world's GDP (7) and total bribes to public officials are estimated at over US $1 trillion annually (8). Furthermore, recent business scandals such as the Volkswagen emission fabrications and

several exchange rate manipulations in the financial industry have eroded trust in the integrity of the corporate world (5). Formal laws and regulations are often designed to enforce honest behavior, but in many situations honest behavior is difficult or impossible to monitor and individuals therefore face a trade-off between honesty personal material gain.

The conflict between honesty and material self-interest is a central feature of human social life, and honesty is exalted in virtually all world religions and moral value systems. Despite substantial interest in the origins and determinants of honest behavior in biology (9), behavioral sciences (2, 10), and economics (4, 11), little is known about the neural processes that enable humans to resolve conflicts between honesty and personal financial gain. Understanding the neural processes involved in these "costly" displays of honesty could offer important new perspectives on the evolutionary origins and development (9, 12) of honest behavior and may also aid in designing interventions for enhancing lie detection, enforcement of honesty (13), and the treatment of pathological cheating (14).

The neural basis of honesty remains largely unexplored in humans because previous studies have almost exclusively relied on instructed-lying paradigms, which examine *deception ability* rather than dishonest behavior (15). Participants in these studies are explicitly instructed by an experimenter to make untruthful statements and also do not benefit materially from lying. Thus, participants neither genuinely decide to be honest nor face a trade-off between honest behavior and material gain. Other recent studies have used signaling games to study the neural basis of deception (16, 17), but such tasks potentially confound honesty with strategic motives (e.g., if senders believe opponents will do the opposite of what they recommend, then a sender will actually "deceive" the opponent by telling the truth (18)). Only one neuroimaging study has investigated cheating in a setting that involved a moral trade-off between honesty and financial gains (19). In that study, honest behavior correlated with brain activity in a network comprising areas of the right dorsolateral prefrontal cortex

(rDLPFC). However, these correlational findings cannot determine whether heightened neural activity genuinely causes honest behavior or simply reflects a functionally irrelevant by-product of honest behavior.

Here we present direct causal evidence for a neural mechanism that increases honesty by applying transcranial direct current stimulation (tDCS) in 145 subjects confronting a real trade-off between honesty and personal material gain. tDCS is a non-invasive method to modulate neural excitability in healthy humans by applying weak electric currents to the scalp (see (20) and (21)). To exogenously enhance neural excitability, we applied anodal tDCS (N=49) over the rDLPFC region previously identified with neuroimaging (Fig. S1B in (19); Fig. S1). We also measured behavior in two additional groups where we applied tDCS to either decrease neural excitability (cathodal, N=49) or leave it unchanged (sham, N=47). Random assignment to conditions generated three groups that were matched in socio-economic status, cognitive ability, and personality. Moreover, the three stimulation conditions were conducted double-blind, were perceived similarly by the participants, and did not differ reliably in terms of participants' mood, alertness, and calmness. Thus, any effects of the three tDCS interventions on honest behavior cannot be explained in terms of pre-existing group differences or changes in beliefs and emotions (21).

During stimulation, we measured cheating using an incentivized and unobtrusive die-rolling task (10, 11) that has been shown to reliably predict rule violating behavior in real-world settings (22). Subjects were instructed to report the outcomes of ten die rolls using a computer interface. Each roll could result with 50% probability in either a gain of 9 Swiss francs or no change in payoff. Prior to each roll, a computer screen indicated which outcomes would yield the monetary payoff. Given that the participants could earn up to 90 Swiss francs (about US $90 at the time of testing) in this task, they faced a substantial material incentive to over-report the number of successful die rolls. Participants completed the task anonymously

(i.e., outside of the experimenter's view) and thus their outcomes could not be independently verified. Although this paradigm cannot identify die rolls for which any individual participant displays dishonest behavior, the degree of cheating associated with each tDCS intervention can be determined by comparing the mean percentage of reported successful die rolls against the 50% benchmark characteristic of completely honest reporting.

Cheating was substantial in the neurally ineffective sham condition (see Fig. 1A). Compared to the honesty benchmark of 50%, participants reported 68% successful die rolls on average (95% confidence interval: 63%-74%). This implies that cheating occurred in 37% of all responses, assuming that participants never misreported to their disadvantage (21). Figure 1B shows the binomial distribution of successful die rolls expected if everyone behaved honestly and the empirically observed distribution for sham tDCS. 8.5% of the subjects reported successful outcomes for all ten rolls (thereby maximizing their earnings), which is significantly higher than the 0.1% expected under the binomial distribution ($p<0.001$, binomial test). Subjects who claimed nine, eight, and seven successful die rolls were also significantly over-represented ($p<0.001$, $p=0.001$, and $p=0.002$, binomial tests), suggesting that many of them cheated on some but not all possible occasions. Such incomplete cheating is commonly observed in similar paradigms (10, 11).

To test whether enhanced neural excitability promotes honest behavior, we compared the distribution of die rolls in anodal and sham tDCS. Anodal stimulation over the rDLPFC substantially reduced the average percentage of successful die rolls to 58% ($p=0.005$, rank-sum test, see Fig. 1A). This corresponds to an implied cheating rate of 15%, a figure that is nearly 60% lower than that observed in the sham condition. We also no longer found significant over-reporting of nine, eight, and seven successful die rolls in the anodal condition ($p=1.000$, $p=0.168$, and $p=0.369$, binomial tests, see Fig. 2D). However, the fraction of

subjects who reported the maximum outcome of ten successful rolls remained essentially unchanged at 8.2%. This suggests that anodal tDCS predominantly reduced cheating in participants who actually pondered the trade-off between honesty and financial gains, but not in those who were committed to maximizing their payoff.

While anodal tDCS decreased cheating, we did not find opposite behavioral effects of cathodal tDCS (Fig. 2D and C). Participants in the cathodal condition reported 67% successful die rolls (95% confidence interval: 61%-73%), which was not significantly different from the success rate reported in the sham stimulation ($p=0.635$, rank-sum test) but significantly higher than the rate reported for anodal tDCS ($p=0.018$, rank-sum test). There are two plausible explanations for why cathodal tDCS did not increase cheating. First, several studies suggest that cathodal tDCS induces less stable cognitive behavioral effects than anodal tDCS (23). Second, the high cheating rate in the sham condition (which is similar, for example, to the cheating rate in a sample of maximum security prisoners (22)) entails that there was little room for tDCS to further increase incomplete cheating. Thus, cathodal stimulation may not induce transitions from incomplete cheating to the more extreme form of cheating on every possible instance. Regardless, the results of the cathodal condition clearly show that any general side effects of electrical stimulation—such as possible discomfort or distraction—cannot account for the substantial reduction in cheating observed when stimulation polarity was reversed (20).

We explored possible mechanisms for why anodal tDCS increased honesty. Our task was designed such that participants had to trade off financial gain for misreporting against the value they assigned to being honest. The stimulated neural process could therefore have strengthened honesty by either (i) decreasing material self-interest, i.e., the subjective value of money, (ii) enhancing the value placed on honesty, or (iii) perturbing the choice process

that trades off these two conflicting motives. We tested these hypotheses with a series of behavioral tasks administered to our participants while they were under the influence of the stimulation (21).

In order to assess whether anodal tDCS reduced cheating by weakening material self-interest, we employed a dictator game that required participants to split money between themselves and well-known charities. Several studies have documented that participants who behave selfish in dictator games cheat more in other tasks (12). This is also evident in our data: Self-interested behavior in the dictator game (i.e., the amount of money kept) was positively correlated with subjects' earnings from the die-rolling task (Spearman's rho=0.266, p=0.001). However, tDCS did not affect the amount of money kept in the dictator game (p=0.989, Kruskal-Wallis test, Table 1) and controlling for dictator game behavior in a regression analysis did not change the effect of anodal tDCS on honest reporting (21). This finding suggests that the increase in honest behavior caused by anodal tDCS is not due to decreased material self-interest.

To test whether anodal tDCS inhibited cheating by increasing the value placed on honesty, we analyzed participants' moral beliefs under the influence of tDCS. Participants indicated the extent to which they considered misreporting in the die-rolling task to be morally inappropriate. This measure was negatively correlated with report rates in the die-rolling task (Spearman's rho = -0.448, $p<0.001$), confirming that participants who highly valued honesty cheated less. However, tDCS did not affect this measure of moral values (p=0.507, Kruskal-Wallis tests). We also did not find that tDCS influenced participants' beliefs about the appropriateness of various forms of dishonest behavior in everyday life situations (p=0.948, Kruskal-Wallis test). Moreover, controlling for participants' ratings on these measures in a regression analysis did not alter the effect of anodal tDCS on cheating (21). Thus, the

reduction in cheating caused by anodal tDCS does not appear to be due to increased moral valuations of honesty.

We next examined whether tDCS stimulation is involved in resolving the trade-off between honesty and material self-gain. If this were the case, then anodal tDCS should primarily influence individuals who were genuinely conflicted between honesty and material gain. As reported earlier, anodal tDCS indeed reduced incomplete cheating but did not alter the rate of complete cheating (Fig. 1B and D). The latter is presumably associated with low conflict due to the complete dominance of financial over moral concerns. We corroborated this result by further examining the magnitude of the tDCS effect in participants who reported low or high moral conflict associated with cheating (see Fig. 2). To this end, we used the median rating (corresponding to the point of indifference on the Likert scale) of how 'morally inappropriate' participants considered cheating in the die-rolling task to divide subjects into a low- and high-conflict group. For low-conflict participants (n=42), cheating rates were unaffected by anodal compared to sham tDCS (p=0.327, rank-sum test). In contrast, high-conflict subjects (n=54) cheated significantly less in the anodal tDCS than the sham group (p=0.014, rank-sum test, n=54). Remarkably, responses for high-conflict subjects who received anodal tDCS were not statistically different from the 50% honesty benchmark (p=0.920, t-test, n=30). These findings substantiate that tDCS only affected the trade-off between honesty and material self-interest for participants who were in fact conflicted between these two motives.

In light of these findings, the question emerges whether the stimulated neural process is specialized for resolving conflicts between material self-interest and honesty, or whether it reflects a general-purpose mechanism involved in any choice between conflicting response options (24). To answer this question, we examined how tDCS affected behavior in three

control tasks that required choices between monetary payoffs associated with different levels of risk, ambiguity, and temporal delay, respectively. The stimulation did not affect choices on any of these tasks (see columns 2 to 4 in Table 1), and controlling for participants' behavior in these tasks in a regression analysis did not alter the effect of anodal tDCS on cheating (21). Thus, the neural mechanism affected by tDCS does not appear to generally affect choices involving financial trade-offs but rather specifically resolves conflicts between material self-interest and the motivation to behave honestly.

A final question we address is whether anodal tDCS over the rDLPFC also reduces cheating when the beneficiary is another person rather than oneself. Testing for such tDCS effect on prosocial cheating is crucial, as it establishes whether the affected neural mechanism is specific to the conflict between honesty and material self-interest rather than controlling cheating in general (regardless of whether the outcomes benefit oneself or others). This test also addresses potential concerns that anodal tDCS may reduce cheating by biasing participants to opt for a response strategy that is less effortful and complex, as reporting the true or default outcome may be easier than generating false responses to earn money (15). In our design, self-interested and pro-social cheating are matched for cognitive complexity, as both require participants to generate fake responses. To test these accounts, we conducted an additional tDCS experiment with 156 participants (anodal N=78, sham N=78) for which we modified the die-rolling task so that subjects could not earn any money for themselves; instead, their earnings were credited to another anonymous participant. All other aspects of the experimental design and procedure were identical to the previous experiment (21).

In line with previous findings (12, 25), participants undergoing sham tDCS cheated even when the associated gains were assigned to an anonymous recipient (Fig. 3A and B). On average, they reported 61% successful outcomes (confidence interval: 56%, 66%), which

corresponds to an implied cheating rate of 22%. A substantial fraction of subjects therefore cheated for purely prosocial reasons, even though the level of cheating was somewhat less pronounced ($p=0.044$, rank-sum test) than in the main experiment where cheating served participants' own interest. However, as illustrated in Fig. 3A and C, anodal tDCS did not reduce prosocial cheating: 62% of the die rolls were reported as successful, which does not significantly differ from the responses under sham tDCS ($p=0.805$, rank-sum test). Moreover, the negative effect of anodal tDCS on dishonest reporting was significantly stronger in the main experiment than in the prosocial die-rolling task ($p=0.017$, Wald test). This finding suggests that the tDCS-induced enhancements of honesty in the main experiment cannot be explained by differences in cognitive effort associated with cheating, and it further indicates that rDPLFC activity is specifically involved in the resolution of conflicts between honesty and self-interest rather than affecting all forms of dishonest behavior.

Our results demonstrate that neural mechanisms involving the right dorsolateral prefrontal cortex play a causal role in modulating honesty when individuals stand to gain from dishonest behavior. These neural processes are functionally independent from other forms of behavioral trade-offs such as those related to risk (26), ambiguity (27), or delayed rewards (28, 29). Such specialization suggests a dedicated neurobiological process that enables humans to resist the self-interested temptation to cheat, consistent with proposals that complex social structures in primate groups have led to the evolution of neural processes dedicated to the control of social behavior (30). This also concurs with evidence from twin studies suggesting that moral beliefs about dishonesty are partially inherited (31).

While the neural process enhanced by tDCS was clearly functionally specific, it is unlikely that it is restricted to the DLPFC area targeted with the tDCS. The neural processes responsible for the enhanced honesty during anodal tDCS are more likely to reflect functional

interactions in a network of brain areas influenced by the stimulation (19, 32). Irrespective of these considerations, the current demonstration of a dedicated neurobiological basis for honesty may have important implications for jurisdiction and legal systems, in which the biological limits on the responsibility for legal transgressions are intensely debated (33). Moreover, our findings of a malleable neural process that influences honesty may be important for the development of measures to enforce honesty (13). However, our finding that tDCS only enhanced honesty in individuals who experienced a conflict when cheating may prevent establishing such measures for the treatment of pathologies coined by an absence of such conflicts (14).

Figures and Tables:

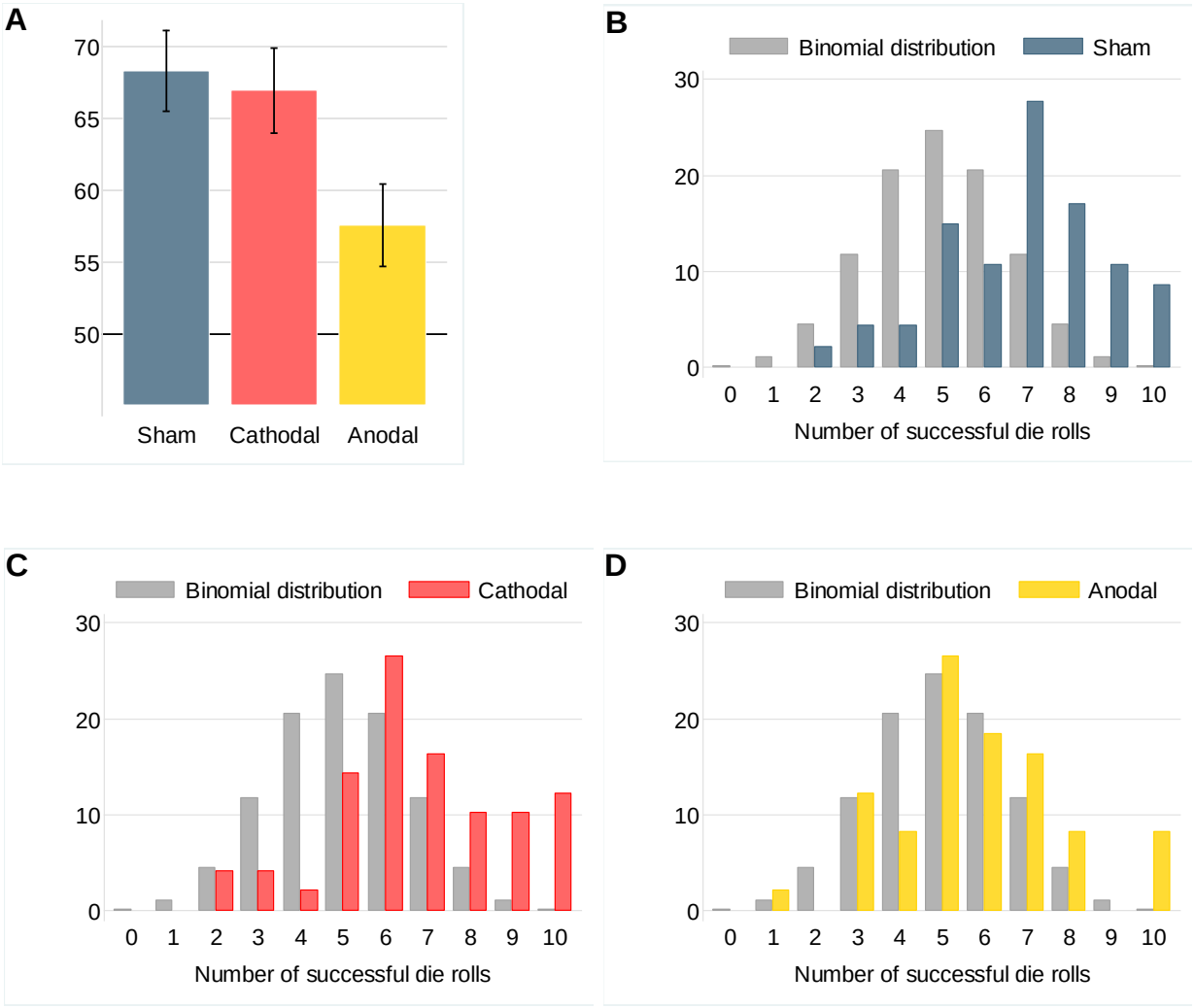


Fig. 1. Effects of tDCS on reporting in the die-rolling task. (**A**) Error bars indicate ±1 SEM. The self-reported percentage of successful die rolls is significantly reduced for anodal (**D**) compared to sham (**B**) and cathodal (**C**) tDCS (p=0.005 and p=0.018, rank-sum tests, n=96 and n=98). The empirical distribution for the cathodal and sham group are skewed towards higher numbers of successful die rolls compared to the binomial (honest) distribution. The distribution for the anodal group more closely resembles the binomial distribution.

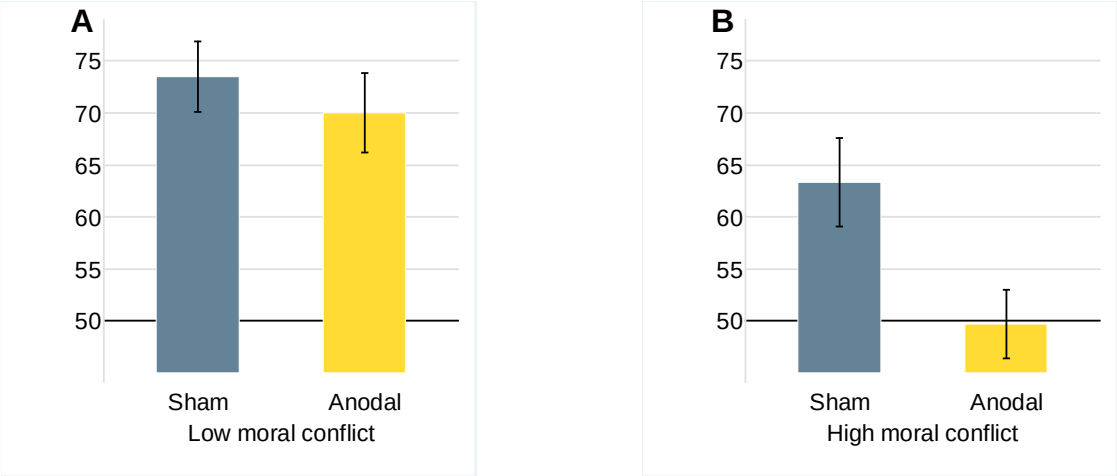


Fig. 2. Effect of anodal tDCS for subjects who experience a low and high moral conflict. (**A**) for subjects who assign a low moral value to honesty (below the median of the group, which corresponds to the point of indifference on the Likert scale), the self-reported percentage of successful die rolls does not significantly differ between anodal tDCS and sham (p=0.327, rank-sum test, n=42). (**B**) For subjects who assign a high moral value to honesty (above or equal to the median), the difference in successful die rolls between sham and anodal tDCS is statistically significant (p=0.014, rank-sum test, n=54). Error bars indicate ± 1 SEM.

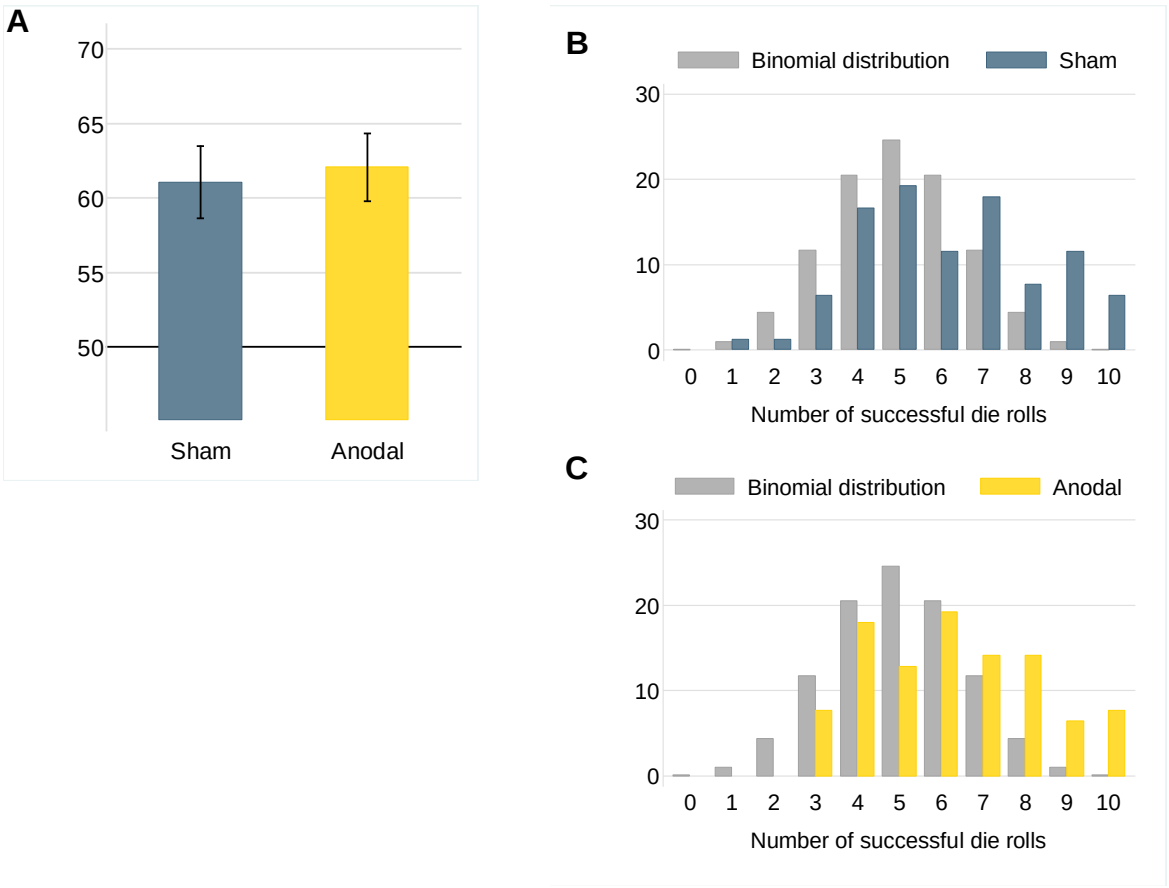


Fig. 3. Effect of anodal tDCS on pro-social cheating. (**A**) Error bars indicate SEM. The self-reported percentage of successful die rolls does not significantly differ between the anodal tDCS and sham (p=0.805, rank-sum test, n=156). The distributions for anodal tDCS (**C**) and sham (**B**) group are similar and both skewed towards higher numbers of successful die rolls compared to the binomial (honest) distribution.

Table 1. Effect of tDCS on other types of behavioral conflicts.

		(1) Self-interest	(2) Risk	(3) Ambiguity	(4) Impulsivity
Sham	Mean	77.433	40.099	29.220	4.716
(N=47)	SEM	3.800	3.438	3.060	0.507
Cathodal	Mean	79.207	44.639	39.864	4.571
(N=49)	SEM	3.295	4.235	3.960	0.553
Anodal	Mean	81.027	38.422	36.435	4.857
(N=49)	SEM	2.626	3.275	3.462	0.509
Kruskal-Wallis	p-value	0.989	0.626	0.139	0.826

Self-interest measures the average percentage of the endowment subjects kept for themselves in the three dictator games. Risk (Ambiguity) is the percentage of the endowment invested into a lottery with known (unknown) outcome probabilities. Impulsivity is the average number of impatient choices (0-20) made in three delay discounting tasks. See methods for details. The Kruskal-Wallis tests in the last row show that tDCS did not have any significant influence on any of these measures of conflict resolution, demonstrating that the stimulated neural process specifically resolves conflicts between honesty and material self-interest. SEM = standard error of the mean.

References and Notes:

1. Colin F. Camerer, Spezio M, Joseph Tao-yi Wang (2010) Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games. *Am Econ Rev* 100(3):984–1007.
2. Irlenbusch B, Villeval MC (2015) Behavioral ethics: how psychology influenced economics and how economics might inform psychology? *Curr Opin Psychol* 6:87–92.
3. Weisel O, Shalvi S (2015) The collaborative roots of corruption. *Proc Natl Acad Sci* 112(34):10651–10656.
4. Gächter S, Schulz JF (2016) Intrinsic honesty and the prevalence of rule violations across societies. *Nature* 531(7595):496–499.
5. Cohn A, Fehr E, Maréchal MA (2014) Business culture and dishonesty in the banking industry. *Nature* 516(7529):86–89.
6. Artavanis N, Morse A, Tsoutsoura M (2016) Measuring Income Tax Evasion using Bank Credit: Evidence from Greece. *Q J Econ*.
7. The Tax Justice Network (2011) The Cost of Tax Abuse: A briefing paper on the cost of tax evasion worldwide. Available at: http://www.taxjustice.net/wp-content/uploads/2014/04/Cost-of-Tax-Abuse-TJN-2011.pdf.
8. World Bank Institute (2004) The Costs of Corruption. Available at: http://go.worldbank.org/LJA29GHA80.
9. Trivers R (2011) *Deceit and Self-Deception: Fooling Yourself the Better to Fool Others* (Allen Lane).
10. Shalvi S, Dana J, Handgraaf MJJ, De Dreu CKW (2011) Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organ Behav Hum Decis Process* 115(2):181–190.
11. Fischbacher U, Föllmi-Heusi F (2013) Lies in Disguise. An experimental study on cheating. *J Eur Econ Assoc*.
12. Maggian V, Villeval MC (2015) Social preferences and lying aversion in children. *Exp Econ*:1–23.
13. Wolpe PR, Foster KR, Langleben DD (2010) Emerging Neurotechnologies for Lie-Detection: Promises and Perils. *Am J Bioeth* 10(10):40–48.
14. Snyder S (1986) Pseudologia fantastica in the borderline patient. *Am J Psychiatry* 143(10):1287–1289.
15. Sip KE, Roepstorff A, McGregor W, Frith CD (2008) Detecting deception: the scope and limits. *Trends Cogn Sci* 12(2):48–53.

16. Zhu L, et al. (2014) Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nat Neurosci* 17(10):1319–1321.
17. Volz KG, Vogeley K, Tittgemeyer M, von Cramon DY, Sutter M (2015) The neural basis of deception in strategic interactions. *Front Behav Neurosci* 9:27.
18. Sutter M (2009) Deception through Telling the Truth?! Experimental Evidence from Individuals and Teams. *Econ J* 119(534):47–60.
19. Greene JD, Paxton JM (2009) Patterns of neural activity associated with honest and dishonest moral decisions. *Proc Natl Acad Sci* 106(30):12506–12511.
20. Woods AJ, et al. (2016) A technical guide to tDCS, and related non-invasive brain stimulation tools. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol* 127(2):1031–1048.
21. Materials and methods are available as supporting information online at the PNAS website.
22. Cohn A, Maréchal MA, Noll T (2015) Bad Boys: How Criminal Identity Salience Affects Rule Violation. *Rev Econ Stud* 82(4):1289–1308.
23. Jacobson L, Koslowsky M, Lavidor M (2012) tDCS polarity effects in motor and cognitive domains: a meta-analytical review. *Exp Brain Res* 216(1):1–10.
24. Fedorenko E, Duncan J, Kanwisher N (2013) Broad domain generality in focal regions of frontal and parietal cortex. *Proc Natl Acad Sci* 110(41):16616–16621.
25. Erat S, Gneezy U (2012) White Lies. *Manag Sci* 58(4):723–733.
26. Mohr PNC, Biele G, Heekeren HR (2010) Neural Processing of Risk. *J Neurosci* 30(19):6613–6619.
27. Huettel SA, Stowe CJ, Gordon EM, Warner BT, Platt ML (2006) Neural signatures of economic preferences for risk and ambiguity. *Neuron* 49(5):765–775.
28. Hare TA, Camerer CF, Rangel A (2009) Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. *Science* 324(5927):646–648.
29. Figner B, et al. (2010) Lateral prefrontal cortex and self-control in intertemporal choice. *Nat Neurosci* 13(5):538–539.
30. Dunbar RIM (2009) The social brain hypothesis and its implications for social evolution. *Ann Hum Biol* 36(5):562–572.
31. Loewen PJ, et al. (2013) The heritability of moral standards for everyday dishonesty. *J Econ Behav Organ* 93:363–366.

32. Driver J, Blankenburg F, Bestmann S, Vanduffel W, Ruff CC (2009) Concurrent brain-stimulation and neuroimaging for studies of cognition. *Trends Cogn Sci* 13(7):319–327.
33. Aspinwall LG, Brown TR, Tabery J (2012) The Double-Edged Sword: Does Biomechanism Increase or Decrease Judges' Sentencing of Psychopaths? *Science* 337(6096):846–849.
34. Bock O, Baetge I, Nicklisch A (2014) hroot: Hamburg Registration and Organization Online Tool. *Eur Econ Rev* 71:117–120.
35. Fischbacher U (2007) z-Tree: Zurich toolbox for ready-made economic experiments. *Exp Econ* 10(2):171–178.
36. Nitsche MA, Paulus W (2000) Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. *J Physiol* 527 Pt 3:633–639.
37. Nitsche MA, Paulus W (2001) Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans. *Neurology* 57(10):1899–1901.
38. Ruff C, Ugazio G, Fehr E (2013) Changing Social Norm Compliance with Noninvasive Brain Stimulation. *Science* 342(6157):482–484.
39. Beharelle AR, Polanía R, Hare TA, Ruff CC (2015) Transcranial Stimulation over Frontopolar Cortex Elucidates the Choice Attributes and Neural Mechanisms Used to Resolve Exploration–Exploitation Trade-Offs. *J Neurosci* 35(43):14544–14556.
40. Nitsche MA, et al. (2007) Shaping the effects of transcranial direct current stimulation of the human motor cortex. *J Neurophysiol* 97(4):3109–3117.
41. Frederick S (2005) Cognitive Reflection and Decision Making. *J Econ Perspect* 19(4):25–42.
42. Steyer R, Schwenkmezger P, Notz P, Eid M (1997) *Der mehrdimensionale Befindlichkeitsfragebogen (MDBF)* (Hogrefe, Göttingen).
43. Houser D, Vetter S, Winter J (2012) Fairness and cheating. *Eur Econ Rev* 56(8):1645–1655.
44. Gneezy U, Potters J (1997) An experiment on risk taking and evaluation periods. *Q J Econ* 112(2):631–645.
45. Cohn A, Engelmann J, Fehr E, Maréchal MA (2015) Evidence for Countercyclical Risk Aversion: An Experiment with Financial Professionals. *Am Econ Rev* 105(2):860–85.
46. Dohmen T, Falk A, Huffman D, Sunde U (2010) Are Risk Aversion and Impatience Related to Cognitive Ability? *Am Econ Rev* 100(3):1238–1260.

47. World Values Survey Wave 6 2010-2014 OFFICIAL AGGREGATE v.20150418. World Values Survey Association (www.worldvaluessurvey.org). Aggregate File Producer: Asep/JDS, Madrid SPAIN.
48. Christie R, Geis FL (2013) *Studies in Machiavellianism* (Academic Press).

Acknowledgments: We thank the staff of the SNS Lab for practical support. The experiments were approved by the Ethics committee of the Canton of Zurich (KEK 2010-0326/3). Financial support from the SNF (CRSII3_141965) and Gottlieb Duttweiler Institute is gratefully acknowledged. The authors report no potential conflicts of interest.

Supporting Information:

Materials and Methods

Figures S1-S2

Tables S1-S8

Supporting Information for

Increasing Honesty in Humans with Electrical Brain Stimulation

Alain Cohn, Michel André Maréchal, Giuseppe Ugazio, and Christian C. Ruff

correspondence to: alain.cohn@chicagobooth.edu, michel.marechal@econ.uzh.ch, or christian.ruff@econ.uzh.ch

This PDF file includes:

Materials and Methods
Figs. S1 to S2

Tables S1 to S8

Materials and Methods

All testing took place in the group testing room of the Laboratory for Social and Neural Systems Research (SNS-Lab) at the University Hospital Zurich. This room contains 14 identical, interconnected computer workstations that are shielded from sight, thereby allowing parallel and anonymous testing of multiple participants. Testing was always conducted in groups of 12 participants unless some participants did not show up. The participants were recruited with the software "h-root" (34) and the experiments were conducted using the computer software z-Tree (35). On average, experimental sessions lasted about 1.5 hours and participants earned about 80 Swiss francs (approximately 82 US dollars at the time of testing). All testing was conducted with 4 experimenters. Two experimenters instructed the participants and mounted the transcranial direct current stimulation (tDCS) electrodes, one experimenter localized the target sites for tDCS, and one experimenter handled all software used for controlling the experimental tasks and tDCS.

1. Main experiment

1.1 Participants

The participants were 145 university students (72 females, mean age 23+/-4, all right-handed) from the Zurich area. All of them were neurologically and psychiatrically healthy, as ascertained by standardized questionnaires, and did not take any medication at the time of testing. All participants gave informed consent and the procedures received ethical approval from the Ethics committee of the Canton of Zurich (see section "6. Human Subjects Approval").

1.2 Transcranical direct current stimulation (tDCS)

In each session, participants were randomly assigned in equal numbers to three tDCS conditions (anodal, cathodal, and sham). The participants in all these conditions were simultaneously stimulated using a multi-channel stimulator (see below) during the experimental tasks. Neither the participants nor the three experimenters interacting with the participants knew which seats were given active or sham stimulation, ensuring that stimulation was administered in a double-blind fashion.

We applied bipolar tDCS by means of a commercially available, CE-certified multi-channel stimulator (NeuroConn, Ilmenau, Germany). This device allows simultaneous stimulation of up to 16 participants with individually-tailored electric currents applied via electrodes mounted on the scalp. tDCS does not directly elicit action potentials, but depending on electrode polarity (i.e., under the anode or the cathode), the applied currents have been shown to either increase or decrease neural excitability, respectively (20, 36). Thus, tDCS is commonly used in cognitive neuroscience studies as a neuromodulator that facilitates or impairs activity elicited by an experimental task (20). We applied this well-established approach and targeted the rDLPFC region that was activated in the study by Greene and Paxton (19) when participants successfully resisted the temptation to lie (see their Figure S1B and Table S2). We applied either anodal tDCS to enhance this activity, cathodal tDCS to weaken this activity, or sham tDCS to control for unspecific stimulation effects. These stimulation protocols were applied in a between subject-design (i.e., three separate groups), as tDCS has long-lasting after-effects (20, 37) that make it difficult to interleave different stimulation protocols within a given session. Moreover, the between-subject design prevented memory or order effects on responding. Each session involved all three tDCS conditions, which were assigned randomly and double-blind to participants, to control for confounds such as effects of task order, experimenter, or time of day.

tDCS was applied by pairs of standard sponge electrodes soaked in saline solution. One of these electrodes (5cm x 7cm) was placed over the rDLPFC region of interest. The other electrode (10cm x 10cm) was placed over the vertex, based on successful stimulation achieved with this electrode montage in previous studies (38, 39). The vertex electrode was chosen to be considerably larger so as to minimize current density and thereby neural stimulation under this electrode (40). tDCS was applied at an intensity of 1.5 mA for 30 minutes (in the anodal and cathodal group) or 60 seconds (in the sham group). The latter condition mimics the tingling sensations at the start of the stimulation but does not provide neurally effective stimulation effects (20). To minimize the sensations at stimulation onset, the current was linearly ramped up (at the start) and down (at the end) over periods of 20 seconds.

To identify the correct scalp sites for electrode placement, we first recruited 43 participants for whom we had already acquired an anatomical MR image (using a Philips Achieva 3T Scanner and a T1-weighted MP-Rage sequence). We then localized in these brain images the site corresponding to the normalized MNI coordinate x,y,z = -26,-53,18 as found in Greene and Paxton (19), as well as the vertex identified by the dorsal confluence of the two principal sulci. For each participant, we determined the scalp coordinates overlying these cortical targets by means of Brainsight 2.0 frameless stereotaxy (Rogue research, Montreal, Canada). We marked for each participant the precise location of these scalp locations in the space of standard EEG 128 surface electrodes caps (Waveguard-Duke cap system, Cephalon A/S, Noerresundby, Denmark, provided by Advanced Neuro Technology, ANT: www.ant-neuro.com). This measurement revealed that the stimulation point on all participants' skulls was localized in an area of 6 cm^2 (see Fig. S1), delimited on the cap by the electrodes RR4-33 (Top-Left), RR3-67 (Top-Right), RA1-63 (Bottom-Left), and RB1-38 (Bottom-Right), superimposing the rDLPFC. For all remaining participants without a T1-weighted structural MRI scan, we therefore localized the center point of the tDCS electrode to lie in the center of these 4 standardized EEG electrode positions as determined by the Waveguard-Duke cap system.

1.3 Experimental tasks and procedures

Pre-stimulation phase

After a brief welcome instruction, participants were allocated to seats in the lab by handing out randomly shuffled cards with seat numbers. While waiting to be called one-by-one to a separate room for stimulation site localization, participants filled out a questionnaire comprising some basic demographic questions (e.g., age, gender, and income) and the Cognitive Reflection Test (CRT) (41). The latter is a simple measure of cognitive skills that takes only a few minutes and is highly correlated with more elaborate measures of cognitive skills, such as the Scholastic Achievement Test (SAT) or the American College Test (ACT). Once all subjects were connected to the tDCS electrodes, we assessed their current state of mood, alertness, and calmness using the multidimensional mood questionnaire (MDBF) (42).

Stimulation phase

The main experiment started with the tDCS stimulation (see section "1.1.2. Transcranial direct current stimulation (tDCS)"). Because the neurophysiological effects of tDCS may take some time to reach stable equilibrium (36), the main experimental tasks (see below) began 3 minutes after the start of the tDCS stimulation. During this time, subjects first answered some questions about their subjective well-being and life satisfaction. Thereafter, subjects were informed via computerized instructions that they would perform four independent tasks for which they could earn money. They were also informed that only one task, selected at random by the computer at the end of the experiment, would be paid out for real. Paying for

one randomly selected task eliminated the possibility for subjects to hedge their income risks across tasks. We randomized the order of the tasks to control for potential spill-over effects. Moreover, as each session involved all three tDCS conditions, the stimulation was fully orthogonal to the order of the tasks. The subjects performed the tasks in full privacy as the research assistants had left the experimental room. If subjects had questions or needed help, they could press a computer key that sent an alarm signal to the research assistants in an adjacent room.
During the tDCS stimulation, subjects performed a total of four experimental tasks:

1) a die-rolling task to measure cheating,
2) a dictator game to measure self-interested behavior,
3) an investment task to measure preferences for risky and ambiguous outcomes, and
4) a delay discounting task to measure impulsivity.

Our task of main interest is the die-rolling task. The other three tasks are auxiliary tasks that allow us to control for general conflict-related choice behavior that has been associated with rDLPFC activity (26, 28, 29, 38) Moreover, we disguised the main purpose of the experiment by embedding the die-rolling task in a larger test battery.

Die-rolling task: The rules of this task required subjects to roll a six-sided die ten times and to report the outcomes of the die rolls. In each round, half of the rolled numbers resulted in a payoff of 9 Swiss francs whereas the remaining numbers yielded no payoff. Prior to each roll, a computer screen displayed which numbers would yield the monetary payoff in that round; the winning numbers changed from round to round. In this task participants faced a real financial incentive to break the rules by misreporting the outcomes of unsuccessful die rolls. Subjects used a cup to roll the die and check the outcomes. They had to report both the outcomes of the die rolls as well as the associated payoffs in order to make sure that everyone understood the consequences of their actions. Because subjects were fully shielded from sight, there was no way anyone (including the experimenters) could detect whether individual subjects misreported the outcomes of their die rolls. However, it is possible to detect cheating at the group level by comparing the mean percentage of successful die rolls reported by the subjects with the 50% benchmark if everyone reported honestly. If we assume that none of the subjects cheated for his or her disadvantage (i.e., by reporting that an outcome is not successful when in fact it is), we are able to calculate cheating rates at the group level (43). Let m be the percentage of misreported rolls. The percentage of outcomes reported as successful p is thus determined by:

$$p = m + (1 - m) \cdot 0.5 = 0.5\,(1 + m).$$

If subjects cheat in a given round, they report a successful die roll outcome with probability 1. However, if they report honestly, they report a successful outcome only with probability 0.5. We can thus characterize the percentage of misreported die rolls by:

$$m = 2 \cdot p - 1.$$

If the random computer draw at the end of the experiment determined that the die-rolling task was selected for payment, we paid out the earnings from all 10 rounds.

Dictator game: In this task subjects could donate money to three well-known charities: the Swiss Red Cross (SRC), the United Nations Children's Fund (UNICEF), and Médecins Sans

Frontières (MSF). For each charity, subjects were endowed with 60 Swiss francs and decided how much of this sum to donate to the respective charity. We use the average amount kept as a measure of the subjects' selfishness. The use of three charities reduces the influence of mismatch between what subjects consider a good cause and the charities' missions. If the dictator game was selected for payment, we paid out one of the three donation decisions, as randomly determined by the computer at the end of the experiment. The main results are robust if we use the decision for each charity separately (see section "3.3. Robustness checks III: Disaggregated behavioral measures").

Investment task: In this task, subjects made two investment decisions (44, 45). In both cases, subjects were endowed with 45 Swiss francs and decided how much to invest in a risky lottery. They could keep the remaining amount for sure. For the first investment choice (the "ambiguity task"), subjects had imperfect information about the success probability of the lottery. They were shown a picture of a plastic box filled with blue, red, and yellow balls in unknown proportion and were told that the computer will draw one ball at random. If the drawn ball was yellow, they won two and a half times their invested amount; for the two other colors, they lost their invested amount. We set the share of yellow, winning balls to 50 percent (unbeknownst to the subjects). For the second investment choice (the "risk task"), the success probability was again 50 percent, but this time subjects knew the success probability. They were told that the computer would draw a random number between 1 and 100; if the chosen number was between 51 and 100, subjects won two and a half times their invested amount, for all other numbers, they lost their investment. The investment amount in the ambiguity task provides us with a measure of subjects' willingness to take risks under ambiguity, which is an inherent feature in most real-life situations that involve risk. By contrast, the investment amount in the risk task measures subjects' risk aversion in the absence of ambiguity. The ambiguity task preceded the risk task in order to prevent that subjects would use the success probability in the risk task as a benchmark for estimating the success probability in the ambiguity task. If the investment task was chosen for payment, one of the two investment decisions became relevant for the payoff, which was randomly determined by the computer at the end of the experiment.

Delay discounting task: In this task, subjects made a series of binary choices between receiving 60 Swiss francs at a later date and obtaining an equal or smaller amounts of money at an earlier date (46). We implemented three scenarios: (i) "today vs. in 3 months", (ii) "today vs. in 6 months", and (iii) "in 3 months vs. in 6 months." For each scenario, subjects were given a list with 20 choice situations between the delayed payment and some earlier payment. In the first row, the amount of the earlier payment was equal to the amount of the delayed payment (i.e., 60 Swiss francs). For every subsequent row, the amount of the earlier payment decreased by 3 Swiss francs. Thus, the sooner subjects switch to the later payment when going through the list, the more patient they are. We used the subjects' average switching point from the earlier to the delayed payment in all three scenarios as a measure of their impulsivity; but note that the results remain robust if we use the switching points of the individual scenarios or a measure of present bias instead (see section "3.3. Robustness checks III: Disaggregated behavioral measures"). If the delay discounting task was chosen for payment, the computer randomly selected one row within a scenario. If subjects chose a payment that was due "today" in that row, they received the money immediately after the testing. If subjects chose the delayed payment (i.e., in 3 months or in 6 months) instead, they could choose between receiving the amount by mail or they could pick it up in person.

Following the four experimental tasks, we again assessed subjects' current state of mood, alertness, and calmness using the MDBF mood questionnaire. This allows us to capture potential changes in mood over the course of the experiment. The mood questionnaire has two versions. We counterbalanced across subjects which version was completed before and after the tDCS stimulation.

Post-stimulation phase

After switching off tDCS, the experiment continued with a final questionnaire. tDCS produces long-lasting physiological after-effects (20, 37), so participants completed the questionnaire while they were still under the influence of tDCS. We first asked participants whether they believed tDCS had an influence on their behavior. We further asked subjects to rate the comprehensibility of the instructions for each of the four experimental tasks. We also assessed subjects' moral valuation of honesty in the context of the die-rolling task, by asking them to indicate the extent to which they agree with the statement "Cheating in the die-rolling task is morally inappropriate" on a 7-point scale ranging from "I do not agree at all" (= 0) to "I totally agree" (= 6). Subjects subsequently answered five questions related to civic honesty as used in the World Values Survey (47). More specifically, the participants had to indicate the extent to which they find certain forms of dishonest behavior in everyday life situations justifiable, such as "Claiming government benefits to which you are not entitled to" or "Avoiding a fare on the public transport", on a 7-point scale ranging from "Never justifiable" (= 0) to "Always justifiable" (= 6). The last part of the survey included a Machiavellianism scale (Mach-IV) (48), which serves as a personality measure of opportunism, status seeking, and lack of morality.

At the end of the experiment the computer randomly selected one of the four experimental tasks for each subject to be paid out. Payoffs were calculated accordingly and paid to the subjects before they left the lab.

Table S1 reports descriptive statistics of the participants from the main experiment. The three experimental groups (i.e., anodal, cathodal, and sham) are well balanced with regard to participants' socio-economic background, cognitive ability, and personality. None of the variables we collected differ statistically between groups at the 5-percent significance level, which demonstrates that the randomization was successful. We nevertheless control for these variables in the regression analysis.

2. Pro-social cheating experiment

We additionally conducted a pro-social cheating experiment in which subjects could not earn any money for themselves in the die-rolling task; instead their earnings from the die rolls were credited to another anonymous participant. All other aspects of the die-rolling task were identical to the main experiment.

The goal of the pro-social cheating experiment was to test whether anodal tDCS over the rDLPFC increases honesty when the financial gains from dishonesty do not serve individuals' self-interest. This allows us to examine whether the stimulated neural process has a specific function to resolve conflicts between honesty and self-interest or whether it is involved in regulating cheating per se (i.e., independent of the underlying motives). This test also addresses potential concerns that anodal tDCS may reduce cheating by biasing participants to opt for a response strategy that is less effortful and complex, as reporting the true or default outcome may be easier than generating false responses to earn money (15). In our design,

self-interested and pro-social cheating are matched for cognitive complexity, as both require participants to generate fake responses.
A total of 156 new subjects were recruited for this experiment and were randomly assigned to anodal and sham tDCS (n = 78 in anodal and n = 78 in sham). All administered procedures and other experimental tasks (including the questionnaires) were identical with those used for the main experiment.
Table S2 provides an overview of the sample characteristics from the pro-social cheating experiment. The two experimental groups (i.e., anodal and sham) are well-balanced in terms of participants' socio-economic background, cognitive ability, and personality. None of the acquired variables differ statistically between groups at the 5-percent significance level, suggesting that the randomization was successful. We nevertheless control for these variables in the regression analysis.

3. Statistical analysis
All reported p-values in the manuscript and supplementary materials are from two-sided tests.

3.1 Regression analysis of the main experiment
Table S3 presents Probit regression results from the main experiment. We estimate the following regression model:

$$\Pr(\text{successful outcome}_{ik} = 1) = \Phi(\alpha + \beta * \text{Anodal}_i + \gamma * \text{Cathodal}_i + \delta * \mathbf{X}_i + \epsilon_{ik}).$$

The decision of individual i to report a successful outcome for die roll k is regressed on indicators for the anodal and cathodal tDCS condition, and a set of control variables $\mathbf{X}_i$. The estimation results remain qualitatively the same if we alternatively estimate a linear probability model using ordinary least squares (OLS).

Column (a) of Table S3 reports the regression results without the control variables. In comparison with treatment sham, anodal tDCS significantly reduced the probability of reporting a successful die roll by 10.6 percentage points (p = 0.007, Wald test). By contrast, the coefficient for cathodal tDCS is close to zero and statistically insignificant (p = 0.737, Wald test). The test results reported at the bottom of column (a) in Table S3 highlights that the coefficient estimates for anodal and cathodal tDCS are significantly different from each other (p = 0.021, Wald test). In column (b), we additionally include a rich set of control variables: age, gender, relationship status, Swiss nationality, monthly amount of cash available, cognitive skills, machiavellism, major of studies, and baseline mood, wakefulness, and calmness. The results remain practically unchanged in comparison to the unconditional regression model.

In column (c) we extend the unconditional model by including a measure of selfishness (i.e., the average amount kept in the three dictator decisions). The results reveal that selfishness is indeed positively correlated with the probability of reporting a successful die roll (p = 0.000, Wald test). However, the coefficient estimate for anodal tDCS remains largely unchanged, suggesting that anodal tDCS did not increase honesty by weakening material self-interest.

Column (d) includes control variables for participants' task-specific valuation of honesty and their perception of civic honesty norms (i.e., their beliefs about the inappropriateness of various forms of dishonest behavior in everyday life situations) (47). While participants who consider misreporting in the die-rolling task to be morally inappropriate were significantly

less likely to report a successful outcome (p = 0.000, Wald test), controlling for their beliefs about the inappropriateness of dishonesty in everyday life did not have any additional explanatory power (p = 0.618, Wald test). The stimulation-induced increase in honest behavior is independent of moral beliefs, as indicated by the relatively stable coefficient of anodal tDCS.

We further estimated a model where we control for measures of general conflict-related behavioral control (i.e., investments in the ambiguity and risk task as well as impulsivity in the delay discounting task). Column (e) shows that none of the three measures are significantly related to subjects' behavior in the die-rolling task (p = 0.614, p = 0.841, respectively p = 0.905, Wald tests) and the inclusion of these measures as control variables in the regression model leaves the impact of anodal tDCS unchanged. Finally, column (f) shows that the effect of anodal tDCS does not change if we include the full set of control variables simultaneously.

3.2 Regression analysis: Comparing selfish and pro-social cheating

In column (a) of Table S4, we perform a difference-in-differences regression analysis to compare the effect of anodal tDCS in the main (selfish cheating) experiment (n = 145) and the pro-social cheating experiment (n = 156). We pool the data from the two experiments and estimate a Probit model in which we include an interaction term between the anodal tDCS condition and a dummy for the main experiment measuring self-interested cheating. The coefficient of anodal tDCS thus captures the treatment effect in the pro-social cheating experiment, while the interaction term shows whether and to what extent the treatment effect differs in the self-interested cheating experiment. We include the same set of control variables used in column (b) of Table S3.

The coefficient of anodal tDCS is close to zero and statistically insignificant (p = 0.757, Wald test), meaning that anodal tDCS did not influence the responses in the pro-social cheating task. This finding contrasts with the treatment effect found in the self-interested cheating experiment, as confirmed by the significant negative interaction effect in columns (a) and (b) of Table S4 (p = 0.022, respectively p = 0.017, Wald tests). Thus, the stimulated neural process is specifically involved in the resolution of conflicts between honesty and self-interest rather than in an inhibition of cheating per se. Moreover, due to the fact that the self-interested and pro-social cheating experiments were equivalent in terms of choice formats and cognitive complexity, these results also establish that the tDCS-induced enhancement of honesty in the self-interested cheating experiment cannot be explained by changes in cognitive effort cost or demands imposed by the cognitive complexity of cheating.

4. Robustness checks

4.1 Robustness checks I: Perception of tDCS and affective state

We examined whether the three tDCS conditions were perceived in a similar fashion by the subjects. Towards the end of the experiment, subjects were asked whether they thought the brain stimulation with tDCS influenced their behavior. Responses to this question did not significantly differ between treatments ($\chi^2 = 2.225$, p = 0.329, χ^2 test), confirming the double-blind nature of the stimulation. Moreover, the regression results reported in columns (a) and

(b) of Table S5 show that including how subjects perceived the stimulation does not alter the effect of anodal tDCS on behavior in the die-rolling task.

To rule out that tDCS may have changed behavior indirectly, by means of altering participants' general affective state, we additionally tested for nonspecific effects of tDCS on subjects' mood, wakefulness, and calmness, by administering a validated questionnaire (MDBF) (42) shortly before and towards the end of the stimulation. tDCS did not influence any of the three mood variables (wakefulness: $\chi^2 = 1.963$, $p = 0.375$; calmness: $\chi^2 = 3.092$, $p = 0.213$; mood: $\chi^2 = 0.150$, $p = 0.928$; Kruskal-Wallis tests). In our regression analysis, we additionally controlled for changes in mood, wakefulness, and calmness and found that the main treatment effect remains robust (see columns c and d of Table S5).

4.2 Robustness checks II: Cognitive complexity

Anodal tDCS could have reduced cheating by biasing subjects to opt for a cognitively less complex response strategy (as reporting the true or default outcome may be easier than generating a false response). However, as we argue in the main text, the selfish and the pro-social cheating experiments are fully matched in terms of cognitive complexity because both required similarly strategic generation of fake responses. If cognitive complexity was responsible for our results, then we should have observed similar effects of anodal tDCS in both experiments. This is clearly refuted by the data.

We nevertheless conducted two further robustness checks to address the issue of cognitive complexity. First, we assessed whether tDCS influenced subjects' understanding of the die-rolling task. At the end of the experiment, we asked subjects to indicate how much they agreed with the statement "the instructions of the die-rolling task were comprehensible" on a scale from "I do not agree at all" (= 0) to "I totally agree" (= 6). The average score was above 5.7 in all three treatments and did not differ between the tDCS conditions ($\chi^2 = 1.114$, $p = 0.573$; Kruskal-Wallis tests). Moreover, the regression analysis in column (b) of Table S6 shows that comprehension of the instructions neither correlates with response behavior in the die-rolling task ($p = 0.764$, Wald test) nor changes the impact of anodal tDCS on cheating. Second, we tested whether cognitive skills are related to behavior in the die-rolling task. The questionnaire from the pre-stimulation phase included the Cognitive Reflection Test (41), a validated measure of cognitive skills. The results reported in column (c) of Table S6 highlight that cognitive skills do not significantly predict behavior in the die-rolling task ($p = 0.132$, Wald test). Additional regression results (not reported in detail) indicate that the interaction effects between cognitive skills and the anodal (or cathodal) tDCS conditions are insignificant ($p = 0.216$, respectively $p = 0.439$, Wald test). In sum, we find no indication that cognitive complexity underlies the effect of anodal tDCS on honest reporting.

4.3 Robustness checks III: Disaggregated behavioral measures

In our main analysis, we used the average amount kept from three dictator decisions as a measure of selfish behavior. The Cronbach's alpha for the three decisions is 0.926, suggesting high internal consistency. In Table S7, we show that tDCS did not have any influence on each of the dictator decisions separately.

Furthermore, in the main analysis we use the average of the switching points elicited in the three delay discounting scenarios (i.e., "today vs. in 3 months", "today vs. in 6 months" and "in 3 months vs. in 6 months") as a measure of impulsivity. Table S8 shows that there is no significant tDCS influence for any of three measures. Moreover, we also constructed a measure of present bias, by taking the difference in switching points between "today vs. in 3

months" and "in 3 months vs. in 6 months". tDCS did not significantly affect this alternative measure of impulsivity either, as shown in the last column of Table S8.

5. Experimental instructions

5.1 Instructions for the main experiment

Below are the instructions for the four experimental tasks from the main experiment. We randomized the order of the tasks at the session level. Subjects were informed that only one of the four tasks was paid out which was randomly determined by the computer at the end of the experiment. We used the term "tokens" to refer to the experimental currency (exchange rate: 100 tokens = 30 Swiss francs).

Task A: Die-rolling task

This part consists of 10 rounds. In each round you can earn up to 30 tokens. The profits from each round will be added up.

In each round, you will be asked to roll a die and report the outcome of the die roll. Use the die and cup in front of you. The die number you report determines your payoff in that round. You can check how much you earn in the payoff table at the bottom of the screen.

Example for the table below: If you roll a 3, your income increases by 0 tokens. If you roll a 4, your income increases by 30 tokens etc.

Payment:

If task A is selected for payment, you will receive the total profit from all 10 rounds.

Please click the OK button to start the task.

Number rolled	1	2	3	4	5	6
Payoff	+0	+30	+0	+30	+30	+0

This screen was shown for all 10 die rolls.

Please complete this task now using the cup at your table.

You can roll the die several rounds to check that it is a fair die. Always remember the first outcome because this is the one that counts.

Number rolled	1	2	3	4	5	6
Payoff	+30	+0	+0	+30	+0	+30

Click the OK button once you have rolled the die.

Please enter now the outcome of the die roll (the first die number you have rolled) and the corresponding payoff.

Number rolled	
Payoff	

Task B: Donation task

In this task you have the opportunity to donate money to three charitable organizations (Swiss Red Cross, UNICEF Switzerland, and Médecins sans Frontières).

For each charity, you get an endowment of 200 tokens. You will have to decide how many tokens you want to donate to each of the three charities. You can keep the remaining amount, i.e., the 200 tokens minus the donation.

If task B is selected for payment, one of your three donations decisions (B1, B2, or B3) will be chosen at random for payment. You can request a copy of the receipts for the overall donation amounts.

Click the continue button if you are ready.
This screen was shown for all three donation decisions.

Your endowment is 200 tokens. You now have to decide what share of your endowment you want to donate to the charity below. You can keep the remaining amount.

[Swiss Red Cross]
[UNICEF Switzerland]
[Médecins sans Frontières]

<u>Your payoff is calculated as follows:</u>

Your payoff = 200 tokens – donation

Please indicate how many tokens you want to donate to [Swiss Red Cross/UNICEF Switzerland/Médecins sans Frontières]: _____ (0-200 tokens)

Task C: Investment task

In this task you will make two investment decisions (C1 and C2) for which you can earn money.

If task C is selected for payment, one of your two investment decisions (C1 or C2) will be chosen at random for payment.

Click the continue button if you are ready.

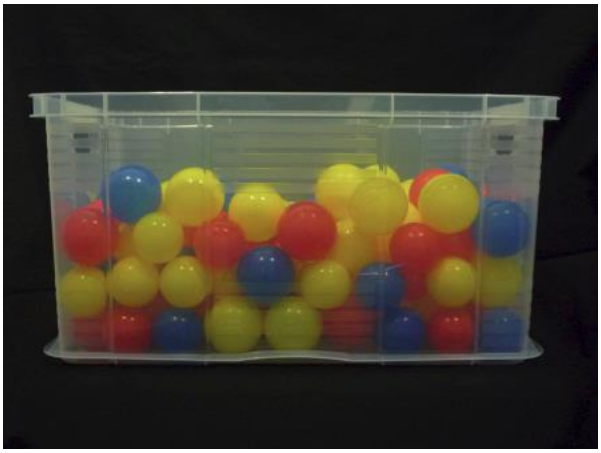
Your endowment is 150 tokens. You now have to decide what share of your endowment you want to invest in a lottery. You can keep the remaining amount that you do not invest.

The investment task works as follows:

The computer will randomly draw one ball out of a box filled with many red, blue, and yellow balls in unknown ratio (see picture below).

- If a red or blue ball is drawn, you will lose your investment and you will not get any money back.
- If a yellow ball is drawn, you win 2.5 times the amount you have invested.

Please note that the proportion of red, blue, and yellow balls is identical to the picture below.



<u>Your payoff is calculated as follows:</u>

- If you lose (a red or blue ball is drawn): Your payoff = 150 tokens – investment
- If you win (yellow ball is drawn): Your payoff = 150 tokens – investment + (2.5x investment)

Please indicate how many tokens you want to invest in this lottery: _____ (0-150 tokens)

The second investment decision works similar as the previous one. Your endowment is again 150 tokens. You now have to decide what share of your endowment you want to invest in a lottery. You can keep the remaining amount that you do not invest.

The investment decision works as follows:

This lottery is different than the previous one. The computer will randomly draw a number between 1 and 100 (each number has the same probability of being drawn).

- If the random number is between 1 and 50, you will lose your investment and you will not get any money back.
- If the random number is between 51 and 100, you win 2.5 times the amount you have invested.

Your payoff is calculated as follows:

- If you lose (random number is between 1 and 50): Your payoff = 150 tokens – investment
- If you win (random number is between 51 and 100): Your payoff = 150 tokens – investment + (2.5x investment)

Please indicate how many tokens you want to invest in this lottery: _____ (0-150 tokens)

Task D: "Sooner vs. later" task

In this task you will have to choose whether to receive a certain amount at an earlier point in time, or whether you prefer to wait in order to get a larger amount at a later time. You will see different decision situations on the screen. For example, "Do you prefer 100 tokens today or 200 tokens in 3 months?" You will have to decide which option you like better.

You will make your decisions based on a total of three choice tables (D1, D2, and D3). In each row, you will see two options, A and B. You can choose between

- option B, a fixed amount of 200 tokens you will get at a later point in time (for example, "in 3 months"),
- or option A, a smaller amount of tokens that will be paid out at an earlier point in time (for example, "today").

If task D is selected for payment, one row from one of the three choice tables will be chosen at random for payment. If you chose the sooner option in that row, you will be paid the according amount at the sooner date. If you chose the later option, you will receive the equivalent of 200 tokens at the later date. If the payment date is "today," you will receive the amount immediately after finishing the study. If the payment date is "in 3 months" or "in 6 months" we will send you the money per mail or you can pick it up at the lab.

Click the continue button if you are ready.

This screen was shown for all three each delay discounting scenarios. We imposed a unique switching point for each scenario. Thus, once participants switched to the delayed payment (i.e., option B), the computer automatically selected the delayed payment for the remaining rows.

Please start with row 1 and then proceed to the next row until you have made a choice in each row. In each row, you have to decide between 200 tokens (option B: "in 3 months") and a smaller amount of tokens (option A: "today"). The amount at the right end of the table (option B) is always the same. Only the amounts on the left side (option A) change from row to row. Make sure to consider the different dates for options A and B. Click the submit button below once you have filled out every row.

Row	Option A	Your choice		Option B
1	200 tokens today	☐	☐	200 tokens in 3 months
2	190 tokens today	☐	☐	
3	180 tokens today	☐	☐	
4	170 tokens today	☐	☐	
5	160 tokens today	☐	☐	
6	150 tokens today	☐	☐	
7	140 tokens today	☐	☐	
8	130 tokens today	☐	☐	
9	120 tokens today	☐	☐	
10	110 tokens today	☐	☐	
11	100 tokens today	☐	☐	
12	90 tokens today	☐	☐	
13	80 tokens today	☐	☐	
14	70 tokens today	☐	☐	
15	60 tokens today	☐	☐	
16	50 tokens today	☐	☐	
17	40 tokens today	☐	☐	
18	30 tokens today	☐	☐	
19	20 tokens today	☐	☐	
20	10 tokens today	☐	☐	

5.2 Instructions for the pro-social cheating experiment

In the pro-social cheating experiment subjects could not earn money for themselves in the die-rolling task; instead they could earn money for another participant.

Task A: Die-rolling task

In this part you will be randomly assigned to another participant who is here today.

This part consists of 10 rounds. In each round you can earn up to 30 tokens for the other participant. The profits from each round will be added up.

In each round, you will be asked to roll a die and report the outcome of the die roll. Use the die and cup in front of you. The die number you report in a round determines the payoff of the other participant in that round. You can check how much you earn for the other participant in the payoff table at the bottom of the screen.

Example for the table below: If you roll a 3, the other participant's income increases by 0 tokens. If you roll a 4, the other participant's income increases by 30 tokens etc.

Payment:

If task A is selected for payment of the other participant, he or she will receive the total profit from all 10 rounds.

The other participants will also perform the die-rolling task and you will be randomly assigned to another participant as well. If this task will be selected for your payment, you will receive the total profit generated by the other participant.

Please click the OK button to start the task.

Number rolled	1	2	3	4	5	6
Payoff for other participant	+0	+30	+0	+30	+30	+0

This screen was shown for all 10 die rolls.

Please complete this task now using the cup at your table.

You can roll the die several rounds to check that it is a fair die. Always remember the first outcome because this is the one that counts.

Number rolled	1	2	3	4	5	6
Payoff for the other participant	+30	+0	+0	+30	+0	+30

Click the OK button once you have rolled the die.

Please enter now the outcome of the die roll (the first die number you have rolled) and the corresponding payoff.

Number rolled	
Payoff for the other participant	

6. Human Subjects Approval

The experiments were approved by the Ethics committee of the Canton of Zurich (KEK 2010-0326/3).

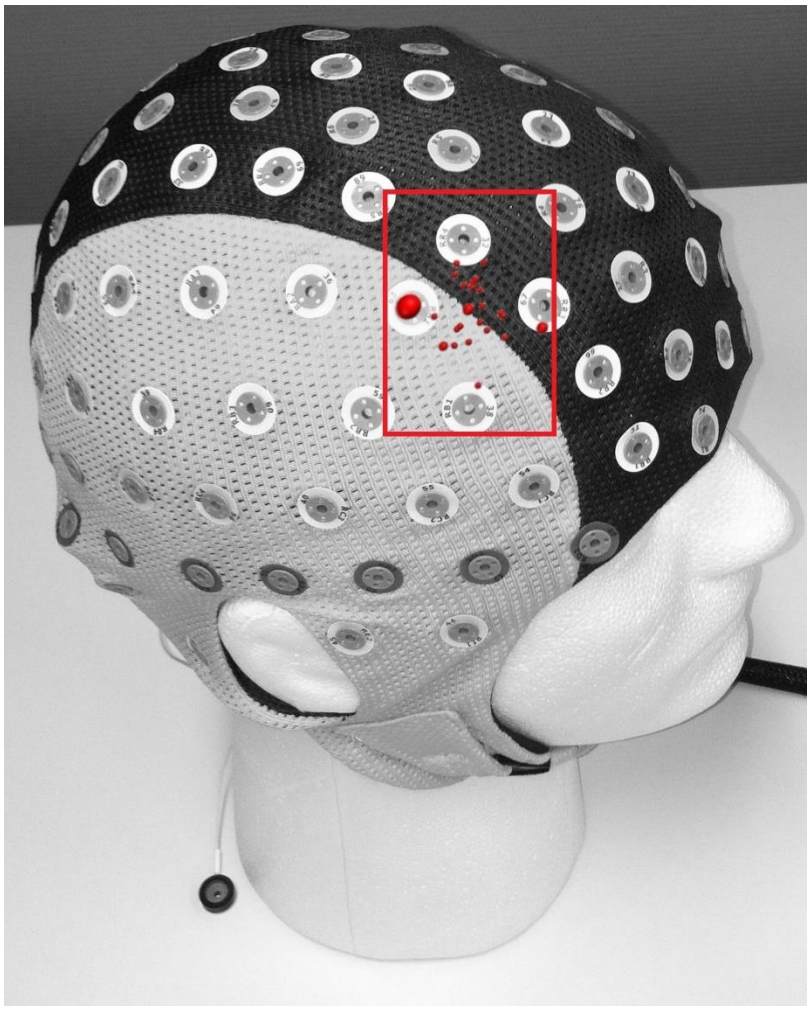


Fig. S1. Schematic Display of the tDCS electrode localization procedure

For the 43 participants with a T1-weighted anatomical MR image, we localized the scalp position overlying the DLPFC coordinate reported in Greene and Paxton (19) to be activated when participants successfully resisted lying. Each red dot represents one such location; the size of the dot represents the number of cases where this location was identified. This illustrates that for all participants, these locations lay in close proximity in an area demarcated by 4 standardized electrode positions in the Waveguard-Duke cap system. We therefore used this cap system in all remaining participants without an MR image to position the tDCS electrode (drawn to real size as red rectangle) so as to cover all possible locations of the DLPFC coordinate.

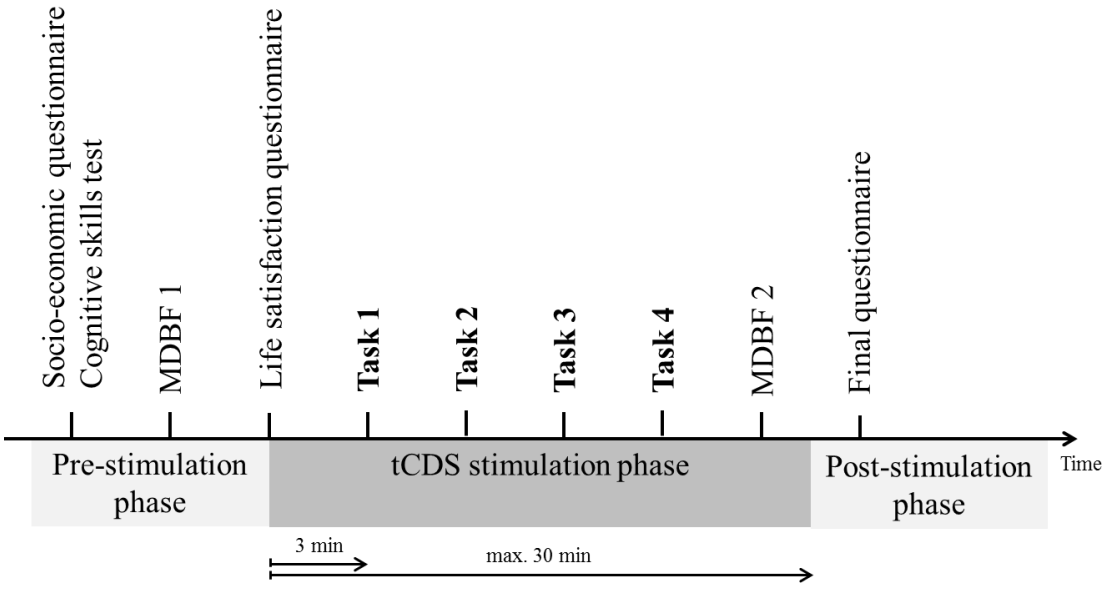


Fig. S2. Timeline of the experiment.

In the pre-stimulation phase subjects completed a socio-economic questionnaire, the Cognitive Reflection Test (CRT) (41), and a mood questionnaire (MDBF) (42). Thereafter, anodal, cathodal or sham tDCS was applied over subjects' rDLPFC. During the stimulation phase, subjects first filled out a life satisfaction questionnaire and then performed four independent tasks (die-rolling task, dictator game, investment task, delay discounting task) in randomized order. In the post-stimulation phase subjects completed another mood questionnaire (MDBF) followed by a final questionnaire.

	Cathodal (N=49)		Sham (N=47)		Anodal (N=49)		Total (N=145)		Kruskal-Wallis / $\chi 2$
	mean	sd	mean	sd	mean	sd	mean	sd	p-value
Age	22.755	3.179	22.957	4.054	22.898	4.547	22.869	3.939	0.865
Male	0.490	0.505	0.553	0.503	0.469	0.504	0.503	0.502	0.694
Single	0.673	0.474	0.723	0.452	0.633	0.487	0.676	0.470	0.636
Monthly income	792.857	551.230	819.149	800.513	605.102	398.858	737.931	606.798	0.181
Swiss	0.755	0.434	0.660	0.479	0.755	0.434	0.724	0.448	0.484
Cognitive skills (CRT)	2.224	0.919	2.553	0.653	2.245	0.969	2.338	0.868	0.190
Machiavellism	-0.227	0.587	-0.305	0.595	-0.201	0.648	-0.243	0.608	0.750
Math	0.429	0.500	0.617	0.491	0.510	0.505	0.517	0.501	0.180
Medicine	0.082	0.277	0.106	0.312	0.061	0.242	0.083	0.276	0.724
Law	0.061	0.242	0.021	0.146	0.061	0.242	0.048	0.215	0.576
Economics	0.041	0.200	0.000	0.000	0.041	0.200	0.028	0.164	0.373
Social sciences	0.224	0.422	0.085	0.282	0.224	0.422	0.179	0.385	0.123
Non student	0.163	0.373	0.170	0.380	0.102	0.306	0.145	0.353	0.576
Positive Mood ($t = 0$)	15.735	2.885	16.468	2.084	16.449	2.102	16.214	2.398	0.514
Wakefulness ($t = 0$)	11.837	2.809	13.468	3.269	12.551	3.594	12.607	3.286	0.062
Calmness ($t = 0$)	15.673	2.726	15.851	2.157	16.449	2.501	15.993	2.482	0.293

Table S1. Descriptive statistics of the individual background variables for the main experiment (n = 145).

The variable 'age' is measured in years; 'single', 'Swiss', and major of study (i.e., 'math', 'medicine', 'law', 'economics', 'social sciences' and 'non student') are binary variables; 'monthly income' is the amount of Swiss francs available to cover living expenses each month; 'cognitive skills (CRT)' is subjects' score (0 to 3) from the Cognitive Reflection Test (CRT) (41); 'machiavellism' is the subjects' score (-3 to +3) on the machiavellism scale (MACH IV) (48); 'positive Mood (t = 0)', 'wakefulness (t = 0)', and 'calmness (t = 0)' are sub-scales (4 to 20) of the multidimensional mood questionnaire (MDBF) (42) administered in the pre-stimulation phase. The last column presents p-values for the null hypothesis of perfect randomization ($\chi 2$ tests in case of binary variables and Kruskal-Wallis tests in case of interval variables).

	Sham (N=78)		Anodal (N=78)		Total (N=156)		Rank-sum / $\chi 2$
	mean	sd	mean	sd	mean	sd	p-value
Age	22.833	3.098	22.987	2.894	22.910	2.989	0.712
Male	0.462	0.502	0.410	0.495	0.436	0.497	0.518
Single	0.679	0.470	0.628	0.486	0.654	0.477	0.501
Monthly income	799.416	579.428	839.744	626.802	819.710	602.111	0.799
Swiss	0.808	0.397	0.769	0.424	0.788	0.410	0.556
Cognitive skills (CRT)	2.333	0.892	2.346	0.880	2.340	0.884	0.953
Machiavellism	-0.252	0.552	-0.210	0.647	-0.231	0.600	0.779
Math	0.487	0.503	0.500	0.503	0.494	0.502	0.873
Medicine	0.115	0.322	0.179	0.386	0.147	0.356	0.259
Law	0.115	0.322	0.077	0.268	0.096	0.296	0.415
Economics	0.000	0.000	0.013	0.113	0.006	0.080	0.316
Social sciences	0.205	0.406	0.179	0.386	0.192	0.395	0.685
Non student	0.077	0.268	0.051	0.222	0.064	0.246	0.513
Positive Mood ($t = 0$)	16.167	2.409	15.795	2.206	15.981	2.310	0.214
Wakefulness ($t = 0$)	12.628	3.520	11.782	3.073	12.205	3.321	0.108
Calmness ($t = 0$)	14.615	2.843	14.744	3.189	14.679	3.012	0.724

Table S2. Descriptive statistics of the individual background variables for the pro-social cheating experiment (n = 156).

The variable 'age' is measured in years; 'single', 'Swiss', and major of study (i.e., 'math', 'medicine', 'law', 'economics', 'social sciences' and 'non student') are binary variables; 'monthly income' is the monthly available amount of cash in Swiss francs; 'cognitive skills (CRT)' is subjects' score (0 to 3) from the Cognitive Reflection Test (CRT) (41); 'machiavellism' is the subjects' score (-3 to +3) on the machiavellism scale (MACH IV) (48); 'positive Mood (t = 0)', 'wakefulness (t = 0)', and 'calmness (t = 0)' are sub-scales (4 to 20) of the multidimensional mood questionnaire (MDBF) (42) administered in the pre-stimulation phase. The last column presents p-values for the null hypothesis of perfect randomization ($\chi 2$ tests in case of binary variables and rank-sum tests in case of interval variables).

Dependent variable:	Successful die roll = 1					
	(a)	(b)	(c)	(d)	(e)	(f)
Anodal	-0.106***	-0.104**	-0.114***	-0.085**	-0.108***	-0.092**
	(0.039)	(0.041)	(0.039)	(0.035)	(0.040)	(0.039)
Cathodal	-0.014	-0.009	-0.017	0.005	-0.019	0.002
	(0.042)	(0.042)	(0.041)	(0.038)	(0.041)	(0.041)
Selfishness			0.002***			0.001
			(0.001)			(0.001)
Value of honesty				-0.045***		-0.042***
				(0.008)		(0.009)
Civic honesty				0.008		0.019
				(0.016)		(0.017)
Ambiguity					0.000	0.000
					(0.001)	(0.001)
Risk					0.000	-0.000
					(0.001)	(0.001)
Impulsivity					-0.001	-0.002
					(0.005)	(0.004)
Additional controls?	No	Yes	No	No	No	Yes
Wald test:						
Anodal = cathodal	0.021	0.011	0.010	0.010	0.025	0.005
Observations	1450	1450	1450	1450	1450	1450

Table S3. Effect of anodal and cathodal tDCS on cheating.

Probit estimates. Reported results are average marginal effects. Robust standard errors, corrected for clustering at the individual level, are displayed in parenthesis. (a) The decision to report a successful die roll is regressed on dummy variables for the anodal and cathodal tDCS treatment in the main experiment (n = 145). (b) The second model includes additional controls for age, gender, relationship status, Swiss nationality, monthly amount of cash available, cognitive skills, machiavellism, major of studies, and baseline mood. (c) The third model includes the average percentage kept in the dictator game as a measure of self-interested behavior. (d) The fourth model includes subjects' task-specific valuation of honesty and their beliefs about the inappropriateness of dishonesty in everyday life. (e) The fifth model includes subjects' investments in the ambiguity and risk task as well as their average impulsivity measured in the delay discounting task. (f) The last model includes all controls simultaneously. The second row from the bottom displays the p-values from Wald tests for the null hypothesis that the coefficients for anodal and cathodal tDCS are equal. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Dependent variable:	Successful die roll = 1	
	(a)	(b)
Anodal X self-interested cheating	-0.118**	-0.122**
	(0.051)	(0.051)
Anodal	0.010	0.013
	(0.033)	(0.031)
Cathodal	-0.014	-0.013
	(0.042)	(0.042)
Self-interested cheating	0.074**	0.066*
	(0.038)	(0.038)
Additional controls?	No	Yes
Observations	3010	3000

Table S4. Effect of anodal tDCS on self-interested and pro-social cheating.

Probit estimates. Reported results are average marginal effects. Robust standard errors, corrected for clustering at the individual level, are displayed in parenthesis. (a) The decision to report a successful die roll is regressed on dummy variables for the anodal and cathodal tDCS treatment in the main experiment (self-interested cheating) and the pro-social cheating experiment (n = 301). We include a dummy for the self-interested cheating experiment and an interaction term between this dummy and the anodal tDCS treatment dummy. (b) The second model includes additional controls for age, gender, relationship status, Swiss nationality, monthly amount of cash available, cognitive skills, machiavellism, major of studies, and baseline mood, wakefulness and calmness. Because one subject failed to respond to some questions, the number of observations drops when adding covariates (n = 300). Significance levels: * p < 0.10, ** p < 0.05, *** p < 0.01.

Dependent variable:	Successful die roll = 1			
	(a)	(b)	(c)	(d)
Anodal	-0.106***	-0.105***	-0.107***	-0.106***
	(0.039)	(0.039)	(0.038)	(0.037)
Cathodal	-0.014	-0.017	-0.006	-0.010
	(0.042)	(0.043)	(0.042)	(0.042)
Belief tDCS influenced behaviour		0.036		0.050
		(0.052)		(0.052)
Δ Positive mood			0.004	0.005
			(0.007)	(0.007)
Δ Calmness			-0.003	-0.003
			(0.006)	(0.006)
Δ Wakefulness			-0.014**	-0.014***
			(0.005)	(0.005)
Wald test:				
Anodal = Cathodal	0.021	0.031	0.010	0.016
Observations	1450	1450	1450	1450

Table S5. Perception of tDCS, affective state, and cheating.

Probit estimates. Reported results are average marginal effects. Robust standard errors, corrected for clustering at the individual level, are displayed in parenthesis. (a) The decision to report a successful die roll is regressed on dummy variables for the anodal and cathodal tDCS treatment in the main experiment (n = 145). (b) The second model includes a dummy for whether subjects believed tDCS influenced their behavior. (c) The third model includes changes in mood, calmness, and wakefulness between the pre- and post-stimulation phase as control variables. (d) The last model includes all controls simultaneously. The second row from the bottom displays the p-values from Wald tests for the null hypothesis that the coefficients for anodal and cathodal tDCS are equal. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Dependent variable:	Successful die roll = 1			
	(a)	(b)	(c)	(d)
Anodal	-0.106***	-0.105***	-0.096**	-0.096**
	(0.039)	(0.039)	(0.040)	(0.040)
Cathodal	-0.014	-0.014	-0.004	-0.004
	(0.042)	(0.042)	(0.042)	(0.042)
Understand instructions		0.009		0.002
		(0.030)		(0.030)
Cognitive skills (CRT)			0.030	0.030
			(0.020)	(0.021)
Wald test:				
Anodal = Cathodal	0.021	0.022	0.018	0.019
Observations	1450	1450	1450	1450

Table S6. Understanding, cognitive skills and cheating.

Probit estimates. Reported results are average marginal effects. Robust standard errors, corrected for clustering at the individual level, are displayed in parenthesis. (a) The decision to report a successful die roll is regressed on dummy variables for the anodal and cathodal tDCS treatment in the main experiment (n=145). (b) The second model includes subjects' understanding of the instructions. (c) The third model includes subjects' score in the Cognitive Reflection Test as a measure of cognitive skills. (d) The last model includes all controls simultaneously. The second row from the bottom displays the p-values from Wald tests for the null hypothesis that the coefficients for anodal and cathodal tDCS are equal. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

		Self-interest I (Red Cross)	Self-interest II (UNICEF)	Self-interest III (MSF)
Sham	Mean	76.830	77.840	77.628
(N=47)	SEM +/-	3.985	4.008	3.849
Cathodal	Mean	80.337	80.051	77.235
(N=49)	SEM +/-	3.401	3.650	3.625
Anodal	Mean	82.235	82.776	78.071
(N=49)	SEM +/-	2.442	2.982	3.290
Kruskal-Wallis	p-value	0.917	0.895	0.854

Table S7. Effect of tDCS on self-interested behavior

Self-interest I, II, and III is the percentage of the endowment that is kept in the dictator game with the Red Cross, UNICEF and Médecins Sans Frontières. The Kruskal-Wallis test in the last row demonstrates that tDCS did not have any significant influence on any of these behavioral measures. SEM = standard error of the mean.

		Impulsivity I (today vs. 3 months)	Impulsivity II (today vs. 6 months)	Impulsivity III (3 months vs. 6 months)	Present bias (Impulsivity III-I)
Sham	Mean	4.957	5.915	3.277	1.681
(N=47)	SEM +/-	0.552	0.675	0.485	0.510
Cathodal	Mean	4.612	5.898	3.204	1.408
(N=49)	SEM +/-	0.620	0.650	0.572	0.526
Anodal	Mean	5.265	6.347	2.959	2.306
(N=49)	SEM +/-	0.664	0.681	0.463	0.619
Kruskal-Wallis	p-value	0.753	0.883	0.736	0.715

Table S8. Effect of tDCS on impulsive behavior

Impulsivity I, II, and III is the average number of impatient choices (0 to 20) in the delay discounting task when subjects had to choose between "today vs. in 3 months," "today vs. in 6 months," and "in 3 months vs. in 6 months." Present bias is the difference in the number of impatient choices between "today vs. in 3 months" and "in 3 months vs. in 6 months." The Kruskal-Wallis test in the last row demonstrates that tDCS did not have any significant influence on any of these behavioral measures. SEM = standard error of the mean.

C. Appendix to Study 3

Brain network mechanisms underlying modulations of social norm complaint behavior by transcranial direct current stimulation

Abbreviated title:

Causal neural networks underlying social norm compliance

Authors and affiliations:

Marius Moisa[1,*,¶], Giuseppe Ugazio[1,2*,¶], Marcus Grueschow[1], Christopher Hill[1], Ernst Fehr[1] and Christian C. Ruff[1,¶]

[1]Zurich Center for Neuroeconomics (ZNE), Department of Economics, University of Zurich, Switzerland

[2] Moral Psychology Research Lab, Department of Psychology, Harvard University Cambridge, USA.

[*] These authors contributed equally to this work.

[¶] **Corresponding authors:**

- Marius Moisa, University Hospital Zurich, Laboratory for Social and Neural Systems Research (SNS Lab), 8091 Zurich, Rämistrasse 100,
  Tel: +41-442551147,
  e-mail: marius.moisa@econ.uzh.ch
- Giuseppe Ugazio, Department of Economics, University of Zurich, 8006 Zurich, Blümlisalpstrasse 10,
  Tel: +41-446345595,
  e-mail: giuseppe.ugazio@econ.uzh.ch
- Christian Ruff, Department of Economics, University of Zurich, 8006 Zurich, Blümlisalpstrasse 10,
  Tel: +41-446345067,
  e-mail: christian.ruff@econ.uzh.ch

Key words:
social norm compliance, combined tDCS-fMRI

Conflict of interest:
The authors declare no competing financial interests.

Acknowledgements:
This work was supported by grants of University of Zurich **(**Forschungskredit, k-33153-01-01) and Zürcher Universitätsverein (Research Talent Development Fund) to M.M. and by grants of Swiss National Science Foundation (105314_152891, CRSII3_141965, and 51NF40_144609) and the ERC ("BRAINCODES") to C.C.R. We thank Karl Treiber for scanning assistance and Cornelia Schnyder for recruiting the participants. All authors gratefully acknowledge support by the Neuroscience Center Zurich (ZNZ).

Abstract

Recent brain stimulation studies have suggested that the right lateral prefrontal cortex (rLPFC) plays a key role in sanction-induced norm compliance (Ruff et al., 2013; Strang et al., 2015); however, the precise neural mechanisms by which stimulation of rLPFC affects norm-compliant behavior remain unknown. Here we investigated this issue by applying anodal-, cathodal- or sham-tDCS over rLPFC during concurrent fMRI of a task measuring norm-compliance triggered by sanction threats (punishment condition) or purely endogenously (no punishment condition). In line with previous results (Ruff et al., 2013), modulating rLPFC activity increased (anodal-tDCS) or decreased (cathodal-tDCS) sanction-induced norm compliance. Importantly, our results indicate that these tDCS-mediated changes in the behavioral response to sanction threats are accompanied by corresponding changes in neural responses to sanction threats in two distinct brain networks: Cathodal-tDCS (which led to weaker sanction-induced compliance) weakened the neural response to punishment threats in several regions within an executive neural network (anterior cingulate cortex [ACC], bilateral LPFC and left parietal cortex). Interestingly, anodal-tDCS (that led to increased sanction-induced norm compliance) did not affect neural responsiveness of this executive network, but rather resulted in stronger amygdala responses to punishment threats as well as augmented punishment-induced functional connectivity between stimulated rLPFC and OFC, and between amygdala and OFC. Thus, our results suggest that the behavioral impact of rLPFC-tDCS on sanction-induced norm compliance may reflect modulations of neural processes involved in both strategy-related and emotional responses to the punishment threat.

Significance Statement

Recent studies suggests that the human brain has developed distinct neural mechanisms that mediate norm-compliant social behavior in the presence of punishment threats. For example, it was shown that lateral prefrontal cortex plays a key role for social sanction-induced norm compliance. Here we combine transcranial direct current stimulation (tDCS) with functional imaging to identify the neural mechanisms that accompany modulation of norm-compliant behavior triggered by anodal-/cathodal-tDCS over lateral prefrontal cortex. We show that decreased norm compliance due to cathodal-tDCS triggers punishment-related activity changes in brain areas implementing strategic thought, while increased norm compliance due to anodal-tDCS is accompanied by increased punishment-related neural responses in brain structures devoted to affective processing. More generally, we show that combined tDCS and functional imaging can identify the causal interplay between behavior, the stimulated site and different neural networks in the service of decision-making.

Introduction

Humans are unique in the extent to which they regulate social life through compliance with social norms (Fehr and Gachter, 2002; Glimcher et al., 2005). Prosocial behavior and widespread cooperation between individuals is made possible by the ability to establish social norms (Hsu et al., 2005). Despite the universal acceptance of such norms, there are always some individuals whose self-interest tempts them to violate the norms. Thus punishments are decisive for norm enforcement and for the maintenance of social order. Since punishment by peers has played an important role in evolution of human society (Smuts, 1999), it is thought that the human brain has developed distinct neural mechanisms that mediate norm-compliant behavior in the presence of punishment (Raine and Yang, 2006; Montague and Lohrenz, 2007; Spitzer et al., 2007; Buckholtz and Marois, 2012). For example, a previous fMRI study (Spitzer et al., 2007) measured brain activity in an economic paradigm specifically designed to test the effect of punishment threats on people's compliance with the fairness norm. The results showed that a fronto-striatal network, including right lateral prefrontal cortex (rLPFC), was activated by the punishment threat. Using the same experimental paradigm, Ruff and colleagues (Ruff et al., 2013) revealed a causal role of rLPFC for social norm compliance by means of transcranial direct current stimulation (tDCS) (Nitsche et al., 2008; Polania et al., 2018). The norm-compliant behavior was increased/decreased when anodal-/cathodal-tDCS was applied over rLPFC. Similar to the cathodal-tDCS effect, a second brain stimulation study also revealed that disrupting the same rLPFC activity by means of transcranial magnetic stimulation decreases the norm complaint behavior (Strang et al., 2015).

However, the existing studies leave it unclear by which neural mechanisms rLPFC-tDCS can affect norm compliance. That is, the behavioral brain stimulation results leave it unclear whether the rLPFC implements only local neural processes which generate the appropriate decisions or whether rLPFC is coordinating the activity in other interconnected brain networks, which in the end are jointly responsible for the change of the norm-compliant behavior. In support for the latter assumption comes the well-established role of the LPFC to coordinate activity in other interconnected brain areas to instantiate behavioral control and action selection of complex processes (Cummings, 1995; Miller and Cohen, 2001; Levine, 2009; Duncan, 2010). Furthermore, it has been proposed that LPFC has an integration-and-selection role in the service of norm complaint behavior and that LPFC integrates and coordinates information from several regions such as amygdala, mPFC or parietal cortex

(Buckholtz and Marois, 2012; Buckholtz et al., 2015). Thus, our hypothesis is that the modulation of the sanction-induced norm-compliant behavior by rLPFC-tDCS is also reflected in sanction-induced activity changes in one out of two distinct networks. On one hand, rLPFC-tDCS might trigger changes in strategic responses to the punishment threat, and thus activity modulation changes in regions within executive network such as parietal cortex, ACC or LPFC are expected. On the other hand rLPFC-tDCS might impact on the affective component of the punishment threat. Thus it might be that the modulation of social norm-compliant behavior by means of rLPFC-tDCS is also drawing on activity modulation of the affective network, namely in regions such as amygdala or orbital frontal cortex.

To test which one of the two hypothesized networks exhibit activity changes with the tDCS modulation of norm complaint behavior, we used tDCS over rLPFC while monitoring the stimulation's impact on both social norm-compliant behavior and on brain activity at the network level by means of functional magnetic resonance imaging (fMRI) (Antal et al., 2011; Moisa et al., 2016). We employed a well-established task (Spitzer et al., 2007; Ruff et al., 2013; Strang et al., 2015), where the participants decide on monetary transfers in an ultimatum game (where the opponent has the possibility to punish if she considers the monetary transfer unfair; punishment condition). As a control, the participants were also performing monetary transfers similar to a dictator game (where punishment is not possible; control condition). Similar to our previous behavioral study (Ruff et al., 2013), we hypothesize that the anodal-/cathodal-tDCS is increasing/decreasing the social norm compliance, as measured in terms of transfer difference between the punishment and the control condition, where no punishment is possible. We also hypothesize that anodal-/ cathodal-tDCS over rLPFC is modulating the punishment-related activity in opposite ways (anodal-tDCS increases and cathodal-tDCS decreases the neural activity), in either regions within executive network or within affective network.

Material and Methods

Subjects

Seventy-nine healthy female volunteers (mean age, 21.56 years; SD, 5.05 years) participated in the experiment. Each participant was assigned to one of the three stimulation groups: anodal-, cathodal- or sham-tDCS. All volunteers provided informed consent to participate and none of them had a history of neurological or psychiatric diseases or used medication regularly. The study was approved by the Research Ethics Committee of the Canton of Zurich.

Experimental task

In the current study we aimed to investigate by which neural mechanisms rLPFC-tDCS can affect social norm compliance. Thus here, we employed a modified version of a well-established experimental paradigm (Spitzer et al., 2007; Ruff et al., 2013) (see Figure 1A). In every round, player A ("the proposer", inside the scanner) was randomly paired with player B ("the responder", outside the scanner) and they interacted anonymously with each other. Player A was endowed with 100 money units (MUs) and proposed a division of these points between both players. Both A and B received 25 MUs extra on every trial, for reasons of fairness and to make social punishment possible. Player A could immediately decide how to distribute the 100 MUs between himself and player B. In the ***control condition*** (no punishment condition), the transfer took place immediately. In contrast, in the ***punishment condition***, player B could spend a part or all of the extra 25 MUs to punish player A if she deemed the split to be unfair. Every MU spent by player B for punishment lead to a reduction of player A's gain by 5 MUs. This means, for example, that if player A transferred nothing to player B, such that after the transfer decision A had 125 MUs and B had 25 MUs, player B could reduce player A's earnings to zero by investing the whole initial endowment for punishing A. Both players were cued whether the current trial was in a control or in a punishment condition. This meant that B was always aware whether she could or could not punish player A if she considered A's choice as violating the fairness norm, and A was always aware of this social punishment threat. All the scanned subjects took the role of "the proposer" (player A) so we could investigate the effect of threat of punishment on their decisions.

Every trial consisted of three phases and the fMRI contrast sensitivity was optimized by simulating optimal task inter trial intervals (Figure 1A). In the first phase (**DECIDE** block in

Figure 1A) a visual cue instructed player A about the current task (control or punishment condition). Once the visual cue was displayed, player A had maximum 4 seconds to decide on the transferred amount. Consecutively player A had to wait for 4.5 +/- a jitter of 2.5 seconds (**WAIT** block in Figure 1A). In the last phase of the trial player A was informed by player B's punishment decision and the associated final payoff (**FEEDBACK** block in Figure 1A). The length of the feedback phase was fixed to 3 seconds. At the beginning of each trial a fixation cross was presented for a random period between 2 to 7 seconds. Thus, the average duration of a trial summed up to 16 seconds (see Figure 1A). In total player A underwent 90 trials, half of which were in the control condition and half with the possibility of being punished. All players A faced the decisions of a randomly selected player B and thus interacted with a real human opponent. All decisions were fully incentive compatible, as the MUs gained by the participants were transformed to Swiss Francs after the experiment according to a predefined conversion rate (1 MU = 0.008 CHF). These earnings were paid out on top of the base pay of 60 CHF.

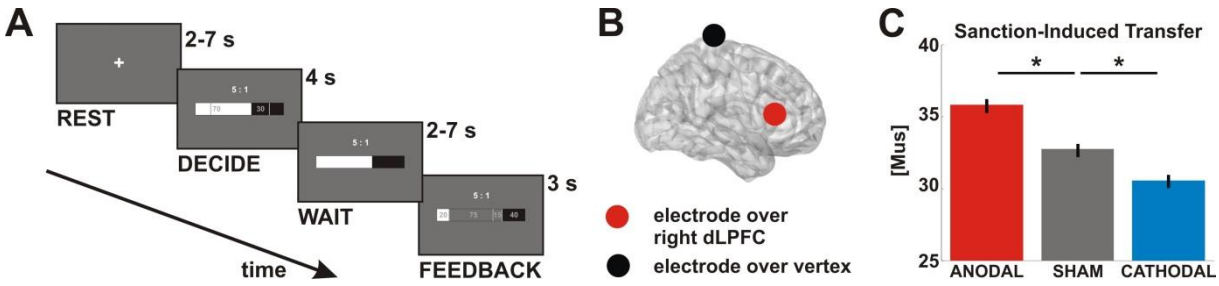


Figure 1 A. Behavioral task. Similar to Spitzer and colleagues (Spitzer et al., 2007), on every trial, player A (inside the scanner) proposed a division of 100 monetary units (MUs) between himself and player B. Both A and B also received 25 MUs extra. In the no-punishment condition (control), the transfer took place immediately, whereas in the punishment condition, player B could spend a part of the extra 25 MUs to punish player A. Every MU spent by player B for punishment led to a reduction of player A's gain by 5 MUs. **B. TDCS setup.** One electrode was placed over rLPFC while the reference electrode was placed over vertex. **C. TDCS impact on behavior**. In line with previous beahvioural findings (Ruff et al., 2013), anodal-tDCS ***increased*** the sanction-induced transfer difference. In contrast, cathodal-tDCS ***decreased*** the sanction-induced transfer difference.

during the fMRI-tDCS sessions but gave their responses in a pilot session recorded beforehand. However, all players B agreed that their responses could be reused in other sessions (Spitzer et al., 2007; Ruff et al., 2013).

TDCS stimulation

For the conccurent combination of tDCS with fMRI we used a previously validated setup (Moisa et al., 2016). For the tDCS stimulation, we used a bipolar MR-compatible current stimulator (DC-Stimulator MC, neuroConn, Ilmenau, Germany) positioned outside the MR-scanner room. TDCS was applied concurrently with fMRI, while the participants were performing the two tasks. For the two active groups (anodal- and cathodal-tDCS) the stimulation lasted for 30 minutes and at the beginning and at the end of the stimulation, the current was ramped up and down over the first and last 30 seconds, respectively. The amplitude of the stimulation was set to 1 mA. For the sham stimulation the current was ramped up to 1 mA over 30 seconds before immediately ramping it down over the next 30 seconds.

Experimental design

Before each participant went into the MR-scanner, the stimulation site over the rLPFC was identified (Figure 1B). We used the exact stimulation montage as in our previous study (Ruff et al., 2013). Thus, the rLPFC was defined using the following MNI coordinates: x=52, y=28, z=14. This standard coordinate was transformed to the individual head-space of each participant using T1-weighted MR scans of participant's neuroanatomy (T1-weighted 3D turbo field echo, 320 sagittal slices, matrix size: 240 x 240, voxel size = 1*1*1 mm, 8-channel MR head coil). The scalp coordinate overlying this brain area was employed as the center point for the active electrode and was determined for each participant prior to the experiment using Brainsight 2.0 frameless stereotaxy (Rogue Research, Montreal, Quebec, Canada). A second reference electrode, (cathode for anodal-tDCS and anode for cathodal-tDCS) was positioned over the vertex, defined in the MR images as the scalp position overlying the confluence of each individual participant's right and left central sulcus. We fixated MR-compatible tDCS electrodes (5 x 7 cm area = 35 cm^2) using a conductive paste (Ten20 EEG Conductive Paste, Weaver and Company, Colorado, USA) over rLPFC and over vertex. Both electrodes were kept in place by means of fixation bandages (DermaPlast CoFix, Hartmann AG, Neuhausen, Switzerland).

Each subject was randomly assigned to one of the three stimulation groups (anodal-, cathodal- or sham-tDCS) and participated in one conccurent tDCS/fMRI experimental session. The acquisition of functional time series started approximately 3 minutes after the start of stimulation, in order to account for possible delays in the onset of stable tDCS effects (Nitsche and Paulus, 2000). Inside the scanner, each participant played the role of player A and underwent 3 experimental runs and a total of 90 trials, half of which were in the control

condition and half with the possibility of being punished (see Experimental task). The order of the trials was pseudo-randomized such that one experimental condition was repeated at most twice in a raw. In total, the task lasted 25 minutes and 30 seconds.

Before the actual start of the stimulation and of the fMRI acquisition, the participants practiced the tasks for 1-2 minutes while lying on the MR-scanner bed. During the practice participants did not get any punishment related feedback in order to avoid possible learning effects.

FMRI acquisition

Functional imaging was performed on a Philips Achieva 3T whole-body MR-scanner equipped with an eight-channel MR head coil. In total we conducted 3 experimental runs, where each contained 250 volumes (voxel size = 3 × 3 × 3 mm^3, 0.5 mm gap, matrix size = 80 × 80, TR/TE = 2100/35 ms, flip angle = 79, parallel imaging factor =1.5, 35 slices acquired in ascending order for full coverage of the brain). We also acquired T1-weighted multi slice fast-field echo B0 scans that were used for correction of possible static distortion produced by the presence of the active electrode (voxel size = 3 × 3 × 3 mm^3, 0.5 mm gap, matrix size = 80 × 80, TR/TE1/TE2 = 481/4.3/7.4 ms, flip angle = 44, no parallel imaging, 37 slices). Additionally, we acquired a high-resolution T1-weighted 3D fast-field echo structural scan used for image registration during post-processing (181 sagittal slices, matrix size = 256 × 256, voxel size = 1 mm^3, TR/TE/TI = 8.0/3.7/181 ms).

Statistical Analyses

Three participants had to be excluded from further analyses. One participant was excluded since she did not understand the task (based on debriefing at the end of the experimental session). Two other participants were excluded since they transferred in all trials (in punishment trials as well as in control trials) 50 MUs. Thus, all the analyses and the results reported here are based on data from 76 participants (24 participants received anodal-tDCS, 26 cathodal-tDCS and 26 received sham-tDCS).

Behavioral analysis

To assess the effects of anodal and cathodal rLPFC-tDCS on the monetary transfer (punishment-induced compared to voluntary norm compliance) we employed a similar analysis as in our previous behavioral study (Ruff et al., 2013). Thus, comprehensive linear mixed-effects (LME) regression analysis was conducted using Matlab. This analysis predicted for each individual *i* the observed choice $T_{i,t}$ in round *t* with the following equation:

$T_{i,t} = \beta_0 + \beta_1(\text{anodal}) + \beta_2(\text{cathodal}) + \beta_3(\text{punishment}) + \beta_4(\text{punishment})*\text{anodal} + \beta_5(\text{punishment})*\text{cathodal} + \varepsilon_{i,t}$ (eq. 1)

Where anodal and cathodal are dummy-coded variables that are set to 1 if individual i received anodal-, cathodal- or sham-tDCS, respectively, or to 0 in all other cases. Punishment is dummy-coded variable that is set to 1 if in the current trial *i* the player B has the option to punish player A, and 0 during the control trials. The model furthermore contained a constant β_0, which measures the average transfer during the control trials during sham-tDCS, as well as the interaction between anodal-tDCS and punishment and the interaction between cathodal-tDCS and punishment. As random effects, the model contained random intercepts for trials and subjects. For completeness, a similar model without random intercepts for subjects and trials revealed a similar statistical result.

FMRI data analysis

Pre-processing of fMRI and GLM design matrix

The fMRI data were analysed with Statistical Parametric Mapping (SPM8, http://www.fil.ion.ucl.ac.uk/spm) implemented in Matlab (MathWorks, Natick, Massachusetts, U.S.A). Pre-processing of the functional time series included motion correction, slice time correction, normalization to Montreal Neurological Institute (MNI) space, spatial resampling to 3 mm isotropic voxels, temporal high-pass filtering and spatial smoothing (Gaussian with 8 mm full-width at half-maximum).

Statistical analysis followed a two-stage procedure. First, we computed a single-subject fixed-effects model for each participant by multiple regression of the voxelwise time series onto a composite model containing the covariates of interest. The GLM design matrix included six main regressors, three per each experimental condition (i.e., DECIDE, WAIT and FEEDBACK for punishment and control condition, respectively). The DECIDE phases were modelled as epochs of duration corresponding to the reaction time of the decision. The WAIT phase and the FEEDBACK phase were modelled as epochs of corresponding lengths (between 2 and 7 seconds for the WAIT block and a fixed duration of 3 seconds for the FEEDBACK block; see Experimental task and Figure 1A). Experimental trials where the participants did not respond were modelled as regressors of no interest.

All regressors were convolved with the canonical hemodynamic response function (HRF). The corresponding temporal and dispersion derivatives of the 6 main regressors were also included in the model. In addition, we also modelled participant-specific head movement parameters to account for BOLD signal changes that correlated with head movements. We

removed possible geometric distortions using the "unwarp" toolbox implemented in SPM8, by means of subject-specific fieldmaps. To allow for group and between-group inferences, we fed the individual contrast images into second-level random-effects analyses. First we investigated which brain regions respond to punishment threats as compared to control (the network that subserves the norm-compliant behavior; for sham group) with a particular focus on regions within the executive and affective network. Consecutively, we focused on identifying regions that exhibit changes in neural responses to punishment threats brought about tDCS (quantified by the interaction between task and type of stimulation, e.g. $[\text{Punishment - Control}]_{\text{Sham}}$ - $[\text{Punishment - Control}]_{\text{Cathodal}}$ or $[\text{Punishment - Control}]_{\text{Anodal}}$ - $[\text{Punishment - Control}]_{\text{Sham}}$; two-sample T tests). As hypothesized, we restricted our search for regions that revealed punishment-related activity modulation brought about tDCS to regions within the executive network or within affective network (see Correction for multiple comparisons section).

We also employed region of interest (ROI) analysis to investigate for possible tDCS modulated activity changes in the stimulated rLPFC. We generated a stimulation site ROI as a 12 mm box cantered on the stimulated rLPFC (MNI coordinates: x=52, y=28, z=14). We extracted parameter estimates (proportional to BOLD signal changes) of the DECISION phase for punishment condition relative to control condition (Punishment - Control). Subsequently we averaged the parameter estimates across all the voxels in the stimulated-rLPFC ROI. We assessed possible stimulation induced modulations of activity in the rLPFC by means of two-sample T-tests (to compare anodal- or cathodal-tDCS groups to sham-tDCS group).

Psychophysiological analysis (PPI)

We also conducted psycho-physiological interaction (PPI) analyses (Friston et al., 1997), to further investigate if any brain regions are showing differential punishment-related functional coupling with the stimulated rLPFC (the contrasts of interest were modelled similarly to the main interactions conducted during the main GLM analysis; e.g. $[\text{Punishment - Control}]_{\text{Sham}}$ - $[\text{Punishment - Control}]_{\text{Cathodal}}$ or $[\text{Punishment - Control}]_{\text{Anodal}}$ - $[\text{Punishment - Control}]_{\text{Sham}}$; two-sample T tests). In this respect, we added to our initial design matrix (1) the BOLD time-series extracted from a 5 mm radius sphere centered around the stimulated rLPFC (MNI coordinates: x=52, y=28, z=14), and (2) the interaction term resulting from the extracted BOLD time-course and all main regressors defined in the initial GLM design matrix (see above), in order to account for the unique effect of the interaction of interest.

The standard GLM analysis revealed punishment-related specific neural modulations brought about anodal-tDCS in a region within the affective network, namely amygdala (see Results section). Thus we conducted a similar PPI analysis, with the seed region in amygdala, in order to investigate if any brain regions exhibit differential punishment-related functional coupling with the amygdala triggered by anodal-tDCS. The seed region was defined as the overlap between the cluster revealed by the main interaction contrast (Punishment - Control]$_{Anodal}$ - [Punishment - Control]$_{Sham}$) and the amygdala ROI within the standard affective network (see Correction for multiple comparisons section).

Correction for multiple comparisons

Our hypothesis was that the two active stimulation conditions will exhibit opposite punishment-related neural changes within executive or affective networks. In this respect we generated the map of the standard executive network by a meta-analysis of 588 studies. Similarly the map of the standard affective network was generated by a meta-analysis of 790 studies (Neurosynth database dated 1st of March 2017, http://neurosynth.org/). The two standard activation maps were corrected for multiple comparisons using an expected false discovery rate of 0.01 and spatially smoothed (Gaussian with 4 mm full-width at half-maximum).

We restricted our analyses to these standard activation maps and we implemented these analyses as non-parametric tests (cluster-level threshold of $p < 0.05$, cluster-forming threshold $T = 2.6$, corrected for multiple comparisons across the emotion or executive network; 5000 permutation, no t-map smoothing) in the software package SnPM (Open source code available at http://warwick.ac.uk/snpm) to optimally correct for type-1 error rates (Eklund et al., 2016). For completeness, we also performed standard SPM whole-brain analyses at a statistical threshold of $p<0.05$ FWE corrected for multiple comparisons at the cluster-level, with cluster-forming threshold $T = 2.6$.

Link between punishment-related neural activity modulation in the stimulated rLPFC and remote brain regions

The fMRI analyses revealed that anodal-/cathodal-tDCS modulates the punishment-related activity in several remote regions both within executive and affective network (see Results sections). Thus, we further investigated the link between the tDCS-related modulation activity in the stimulated site and the punishment-related changes brought about tDCS in these remote brain areas. We performed separate analyses in which we regressed the neural activity in these remote regions affected by tDCS (parameter estimates [PE_remoteRegion],

proportional to BOLD signal changes, for [Punishment – Control] contrast) on both the tDCS intervention (anodal- or cathodal-tDCS) and on parameter estimates of the stimulated rLPFC (for the same contrast, [Punishment – Control]). The models further included the corresponding interactions between the stimulation types (anodal- or cathodal-tDCS) and the parameter estimates in the stimulated rLPFC (PE_rLPFC):

$$PE\_remoteRegion = \beta_0 + \beta_1(anodal) + \beta_2(cathodal) + \beta_3(PE\_rLPFC) + \beta_4(PE\_rLPFC)*anodal + \beta_5(PE\_rLPFC)*cathodal \quad \text{(eq. 2)}$$

where anodal and cathodal were dummy-coded variables that are set to 1 if the current participant received anodal- or cathodal-tDCS, respectively, or to 0 in all other cases. These analyses were performed using the linear mixed effects (LME) function implemented in Matlab.

Results

rLPFC-tDCS effects on social-norm complaint behavior

In line with previous results (Ruff et al., 2013), the behavioral analysis revealed that rLPFC-tDCS changed the social norm compliance, as measured in terms of transfer difference between punishment and control condition, where no punishment is possible. Similar to Ruff and colleagues (Ruff et al., 2013), anodal- and cathodal-tDCS changed sanction-induced norm compliance in opposite ways relative to sham-tDCS. Specifically anodal-tDCS increased the monetary transfer difference (LME analysis, $p < 0.001$; see Figure 1C), whereas cathodal-tDCS decreased the monetary transfer difference (LME analysis, $p = 0.001$; see Figure 1C). Thus, we successfully replicated previous findings of rLPFC-tDCS impact on norm-complaint behavior.

Neural network for social norm compliance: increased punishment-related activity in executive network

Having successfully replicated previous behavioral effects of the stimulation on sanction-induced social norm compliant behavior (Ruff et al., 2013), we proceeded to analyze the fMRI data. The first analysis focused on disclosing the neural network that subserves social norm complaint behavior, computed as the contrast between the brain activations during decision-making epochs under the social punishment threat and activations during decisions in the control condition. Specifically we tested for punishment-related activity changes in executive or affective network for sham-tDCS group. Indeed we observed punishment-related activity increases in regions within executive network such as bilateral dorsal LPFC, left ventral LPFC, bilateral parietal and ACC (Figure 2). No region within the affective network exhibited punishment related increases as compared with control. For completeness, whole brain analysis also revealed increased punishment-related activity in precuneous. No regions exhibited decreased punishment-related activity as compared with control task. Thus, our results revealed punishment-related activations in executive network and no activity modulation of affective network.

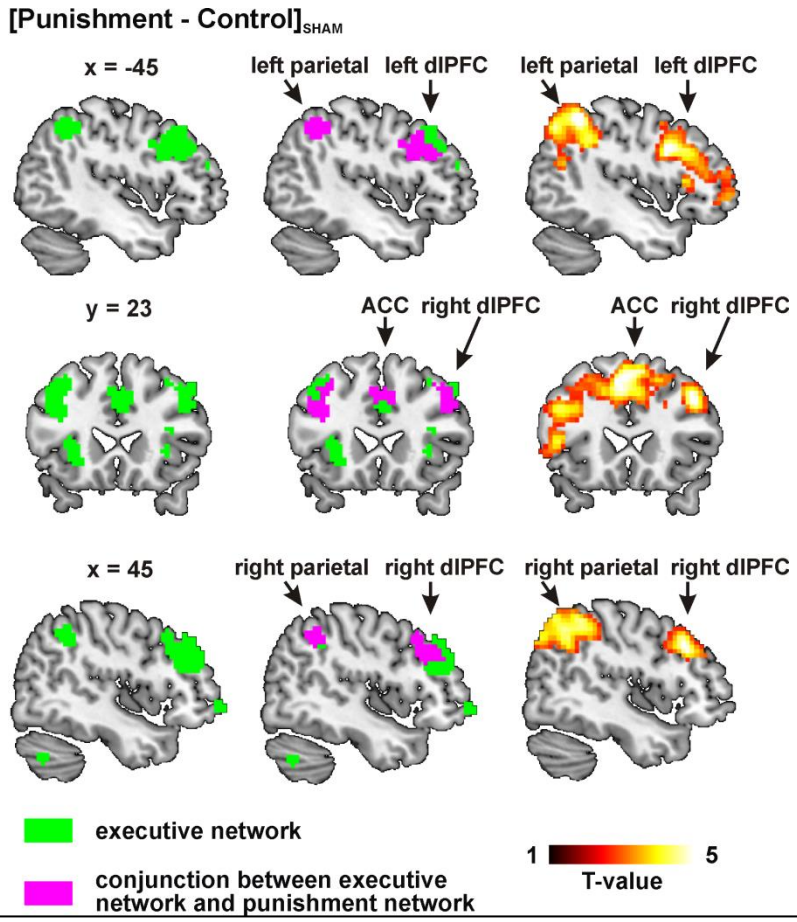


Figure 2 Neural network for social norm compliance. We observed increased brain activity in response to punishment threats compared with control in regions within the executive network, such as bilateral dorsal LPFC (dorsolateral prefrontal cortex), left ventral LPFC (ventrolateral prefrontal cortex), bilateral caudate nucleus bilateral IPL (inferior parietal lobe) and ACC (anterior cingulate cortex). The left brain views display standard ROIs within the executive network, while the right brain views display the regions that revealed punishment-related activity increases as compared with control (standard whole brain SPM analysis; $p<0.05$ FWE cluster corrected, cluster-forming threshold $T > 2.6$). The middle brain views display the overlap between the standard

related activity increases in the executive network. We now turn to the crucial goal of this study, namely to disclose which brain networks exhibit punishment-related dynamic changes brought about rLPFC-tDCS. Is the rLPFC-tDCS modulating the punishment-related activity in the executive network or in the affective network? To this end, we analyzed the individual contrasts between DECISION epochs in trials were punishment was possible versus DECISION epochs in trials were punishment was not available, in second-level analyses

(two-sample t-tests) for different types of stimulation (e.g. cathodal- versus sham-tDCS or anodal- versus sham-tDCS).

During sham-tDCS, these regions strongly responded to punishment threat as compared with control (Figure 3B; post-hoc paired t-test on PEs, $p < 0.05$), whereas during cathodal-tDCS the punishment-related activity blunted out (Figure 3D). Surprisingly, anodal-tDCS did not modulate (increased or decreased) the punishment-related activity in any regions within executive network. Thus, our results hint towards a decrease in strategic response to the punishment threat, effect specific to cathodal-tDCS.

Next, we investigated if rLPFC-tDCS triggers changes in functional connective between the stimulated rLPFC and remote brain regions within executive network. No brain region within the executive network exhibit punishment-related changes (increases or decreases) in functional connectivity with the stimulated rLPFC, neither brought about cathodal- nor by anodal-tDCS.

Anodal-tDCS modulates the punishment-related activity in amygdala (affective network)

Next, we tested for punishment-related modulations in regions within the affective network. Interestingly increased sanction-induced monetary transfer difference by anodal-tDCS was also reflected in punishment-related increased brain activity in amygdala, region within the affective network (Figure 3C & 3D; post-hoc paired t-test on PEs, $p < 0.05$). In contrast, for the sham-tDCS group, the activity in the same amygdala is decreased during the punishment threat compared with control (Figure 3C & 3D; post-hoc paired t-test on PEs, $p < 0.05$). No regions within the affective network revealed any punishment related modulations (increases or decreases) with the cathodal-tDCS. These results suggest that anodal-tDCS brings about an increased emotional response to the punishment threat. Thus, our analyses did not revealed the hypothesized opposite punishment-related activity modulation of regions within the same brain network (for anodal- compared to cathodal-tDCS), but rather a selective modulation of the executive network (due to cathodal-tDCS) and of the affective network (due to anodal-tDCS), respectively.

Cathodal-tDCS modulates the punishment-related activity in the executive network

First we tested for punishment-related modulations brought about rLPFC-tDCS in regions within the executive network. Indeed, decreased sanction-induced monetary transfer difference by cathodal-tDCS compared to sham-tDCS was also accompanied by activity

shifts in regions within executive network (e.g. ACC, bilateral LPFC, left parietal cortex; Figure 3A & 3B).

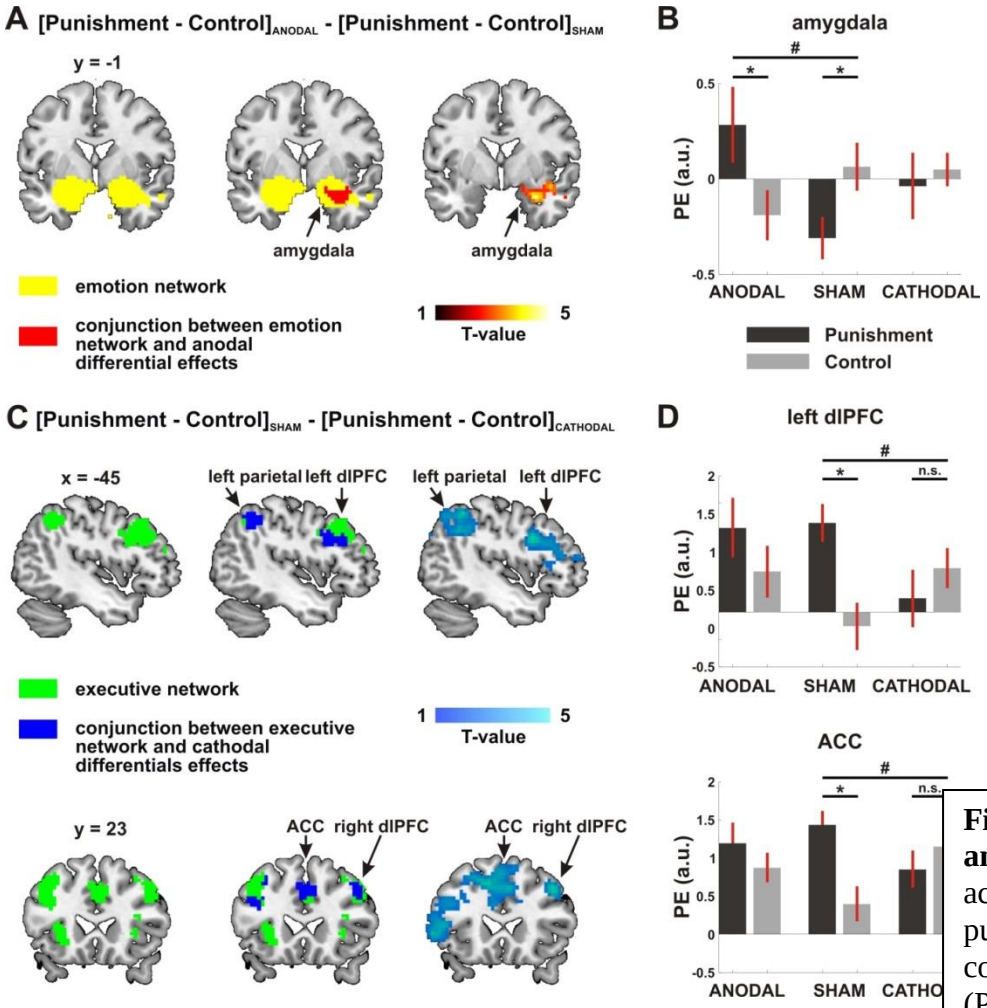


Figure 3 Brain activity changes l and D). A. The ***increased*** sanction accompanied by ***increased*** brain a punishment threats in the amygdal corrected, cluster-forming thresho (PE; proportional to BOLD signal computed with SPM8; * denotes p induced norm compliance due to c during punishment threats compar dorsal LPFC and left parietal corte corrected, cluster-forming thresho (PE; proportional to BOLD signal ACC (# denotes significant intera 0.05).

Anodal-tDCS changes the punishment-related functionally conne stimulated rLPFC and OFC/vmPFC

Next, we investigated if rLPFC-tDCS triggers changes in functional co stimulated rLPFC and remote brain regions within affective network. Indeed, a PPI analysis revealed an increased functional connectivity between the stimulated rLPFC and OFC/vmPFC due to anodal-tDCS (Figure 4A & 4B; post-hoc paired t-test, $p < 0.05$).

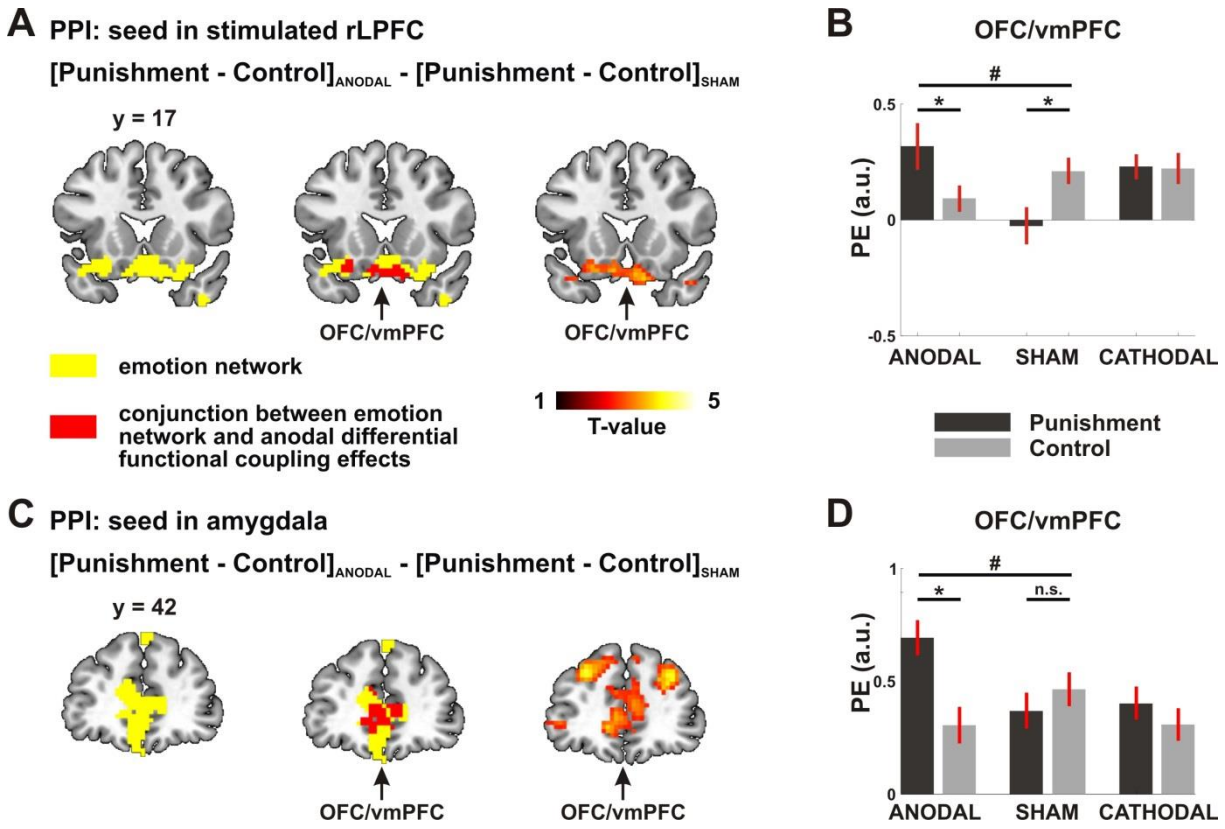


Figure 4 A. Psychophysiological interactions analysis with the seed in the stimulated rLPFC. Anodal-tDCS ***increased*** punished-related connectivity between the stimulated rLPFC and OFC/vmPFC (region within the affective network; standard whole brain SPM analysis; p<0.05 FWE cluster corrected, cluster-forming threshold T > 2.6). **B.** The anodal-tDCS effects on connectivity with stimulated rLPFC, with parameter estimates (PE; proportional to BOLD signal changes) in OFC/vmPFC (# denotes significant interaction computed with SPM8; * denotes post-hoc paired t-test, p < 0.05). **C.** PPI analysis with the seed in amygdala. Anodal-tDCS ***increased*** punished-related connectivity between the amygdala and OFC/vmPFC. **D.** The anodal- tDCS effects on connectivity with amygdala, with parameter estimates (PE; proportional to BOLD signal changes) in OFC/vmPFC (# denotes significant interaction computed with SPM8; * denotes post-hoc paired t-test, p < 0.05).

In contrast, during sham-tDCS, the functional connectivity between rLPFC and OFC/vmPFC decreased during punishment threat compared with control (Figure 4A and B; post-hoc paired t-test, p < 0.05). There were no regions within the affective network showing changes (increases or decreases) in punishment-related functional connectivity with the stimulated rLPFC due to cathodal-tDCS. Thus anodal-tDCS also modulates the connectivity between the stimulated rLPFC and OFC/vmPFC, region within the affective network.

Anodal-tDCS changes the punishment-related functionally connectivity between amygdala and OFC/vmPFC

Our analyses identified two main aspects at the neural level that were affected by anodal-tDCS: first anodal-tDCS increase on social norm-complaint behavior was also accompanied by punishment-related activity increase in amygdala; second anodal-tDCS also increased the punishment-related functional connectivity between the stimulated rLPFC and OFC/vmPFC. To further investigate if anodal-tDCS also modulates the link between these two regions of the affective network (amygdala and OFC/vmPFC) we performed a functional connectivity analysis with the seed region in amygdala. Indeed, this PPI analysis revealed an increased punishment-related functional connectivity between amygdala and OFC during anodal-tDCS (Figure 4C & D; post-hoc paired t-test, $p < 0.05$). In contrast, for sham-tDCS, the functional connectivity between amygdala and OFC decreased during punishment threat compared with control (Figure 4C and D; post-hoc paired t-test, $p < 0.05$). Thus anodal rLPFC-tDCS also modulates the connectivity between these two remote regions within the affective network.

Cathodal stimulation decreases the norm complaint related activity in the stimulated rLPFC

Up to now, the current study successfully confirmed previous rLPFC-tDCS effects on norm complaint behavior (Ruff et al., 2013). The results so far indicate that anodal- and cathodal-tDCS over rLPFC modulation on norm-complaint behavior is also reflected in neural activity changes within two distinct networks, e.g. executive network (due to cathodal-tDCS) and affective network (due to anodal-tDCS). Next we investigated how theses distinct punishment-related effects are mechanistically brought about anodal- and cathodal-tDCS. First we investigated if tDCS over rLPFC induced modulations on punishment-related functional activity at the stimulation site. We compared the activity during the punishment condition relative to the control for anodal- or cathodal-tDCS versus sham-tDCS. Indeed, an ROI analysis (see Methods) of the stimulated area showed that, compared to the sham-tDCS, cathodal-tDCS significantly reduced the rLPFC neural activity associated with punishment-induced norm compliant behavior (rLPFC ROI, two-sample T-test, $P(51) = 0.017$). There was no significant difference in the punishment-induced related activity modulation at the stimulated site for anodal- versus sham-tDCS (rLPFC ROI, two-sample T-test, $P(48) = 0.34$). Thus, our results indicate that the cathodal-tDCS modulatory effects on remote regions within executive network are induced via direct modulation of punishment-related activity in the stimulated rLPFC.

Anodal tDCS increases the correlation between the activity in stimulated rLPFC and the activity in amygdala and OFC/vmPFC

To further investigate the link between the impact of the tDCS on the stimulated rLPFC and the remote impact of the tDCS on regions within the executive or affective network, we performed several regression analyses. First, we regressed the activity change for punishment versus control in amygdala or in OFC/vmPFC, on the neural activity for the same contrast in the stimulated rLPFC and its interaction with the stimulation condition (see methods for details on the regressions). These analyses revealed that, compared to sham-tDCS, only during anodal-tDCS and not during cathodal-tDCS, a stronger punishment-related neural activity change (punishment versus control) in amygdala was associated with a stronger punishment-related neural activity change in the stimulated rLPFC (interaction between neural activity at the stimulation site [PE_rLPFC] and stimulation condition [anodal-tDCS], $t(70) = 1.67$, $p = 0.049$, one tailed; see also Methods section). Similarly, compared with sham-tDCS, during anodal-tDCS a stronger punishment-related neural activity change (punishment versus control) in OFC/vmPFC was associated with a stronger punishment-related neural activity change in the stimulated rLPFC (interaction between neural activity at the stimulation site [PE_rLPFC] and stimulation condition [anodal-tDCS], $t(70) = 2.36$, $p = 0.01$, one tailed; see also Methods section). For completeness, no region within the executive network revealed any significant associations with the punishsment-control related activity in the stimulation site brought about cathodal tDCS. Thus, these results suggest that the pathway by which anodal-tDCS affects subsequent norm-complaint behavior is via a direct link between the stimulated rLPFC with amygdala and OFC/vMPFC (regions within affective network).

Discussion

In this study, we combined concurrently tDCS and fMRI to reveal the brain networks that accompany tDCS-induced changes in social norm-compliant behavior. In line with previous results (Ruff et al., 2013), anodal-/ and cathodal-tDCS increased/decreased norm-compliant behavior, as measured in terms of transfer difference between the punishment and the control condition, where no punishment is possible. Our results indicate that anodal- and cathodal-tDCS over rLPFC triggered punishment-related neural reconfigurations in two distinct brain networks: Cathodal-tDCS (which led to weaker sanction-induced compliance) weakened the neural response to punishment threats in regions within an executive neural network (ACC, bilateral LPFC and left parietal cortex). Interestingly, anodal-tDCS (that led to increased sanction-induced norm compliance) did not affect neural responsiveness of this executive network, but rather resulted in stronger amygdala responses to punishment threats as well as augmented punishment-induced functional connectivity between stimulated rLPFC and OFC, and between amygdala and OFC. Below we discuss all these results in details.

The results obtained in the present study largely corroborated the existing evidence, at both behavioral and neural level. Behaviorally, we replicated the impact of rLPFC-tDCS, confirming that anodal-/cathodal-tDCS rendered participants more/less sensitive to the presence of sanction threats in reference to the control condition, where no punishment is possible, causing larger/smaller adjustments of behavior to the external incentives. At the neural level, we show punishment-related activations (the network that subserves norm-compliant behavior) only in regions within the executive network while the affective network is not modulated by the punishment threat. The punishment-related network revealed here also broadly overlaps with the norm-complaint network described in a previous fMRI study (Spitzer et al., 2007). Thus here we replicate both previous behavioral (Ruff et al., 2013) and imaging results (Spitzer et al., 2007) that investigated the norm-based social behavior.

When examining how tDCS affected the neural activity elicited by the stimulation-related modulation of norm-complaint behavior – the main aim of our study – we did not observed the expected opposite punishment-related activity modulation within a unique network of regions due to anodal- and cathodal-tDCS. Instead anodal- and cathodal-tDCS triggered neural activity reconfigurations in two distinct brain networks: affective network (due to anodal-tDCS) and executive network (due to cathodal-tDCS), respectively. One possible explanation could be that rLPFC is connected and communicates with different networks, such as executive and affective networks in different activity regimes. Previous work has

suggested the core role of the LPFC in coordinating the activity in other interconnected brain areas to instantiate behavioral control and action selection of complex processes (Cummings, 1995; Miller and Cohen, 2001; Levine, 2009; Duncan, 2010). Specifically, it has been proposed that LPFC has an integration-and-selection role in the service of norm-complaint behavior and that LPFC integrates and coordinates information from several regions within both executive and affective network (Buckholtz and Marois, 2012; Buckholtz et al., 2015). In our study cathodal-tDCS resulted in activity changes in regions within the executive brain network (e.g. ACC, bilateral LPFC, left parietal cortex). During sham stimulation, these regions strongly responded to punishment threat as compared to control (broadly overlap with the network that subserves norm complaint behavior; see Figure 2). However cathodal-tDCS blunted-out the punishment-related responses as compared with the control task. These activity shifts of the cathodal-tDCS on regions within the executive network suggest a diminished strategic thinking about the consequences of the punishment threats.

Critically, anodal-tDCS over rLPFC lead to an increase in neural sensitivity to punishment threats in the amygdala, a brain area known to have an important role in processing fear and aversive responses to threats of various nature (Ohman, 2005) and in particular during social interactions (Bechara et al., 2003). Importantly, our results also showed that anodal-tDCS increases the punishment-related functional connectivity both between the stimulated rLPFC and OFC/vmPFC as well as between amygdala and OFC/vmPFC. It is well established that amygdala and the OFC are anatomically and reciprocally connected (McDonald, 1998; Hoover and Vertes, 2007). Also amygdala and the vmPFC are interconnected brain structures that mediate the extinction of conditioned fear both in rats (Milad and Quirk, 2002; Amano et al., 2010) and in humans (Phelps et al., 2004). Furthermore, a whole body of research indicates that neural connectivity between amygdala and OFC is important for updating cue values after changes in their associated outcome (Saddoris et al., 2005; Murray, 2007; Morrison et al., 2011). Here the increased norm-complaint behavior brought about anodal-tDCS is also reflected in changes in the functional interplay between the stimulated rLPFC, amygdala and OFC/vmPFC. Thus, from a functional point of view, our results suggest that the anodal-tDCS may have induced increased affective responses to punishment threats.

The modulation of punishment-related activity in different brain networks (affective network due to anodal-tDCS and executive network due to cathodal-tDCS) might also be accounted by the fact that anodal- and cathodal-tDCS affect different types of neuron populations that might have different connections with regions within different networks. This is only

speculative and future work should further investigate the anodal- and cathodal-tDCS modulatory effects on these two distinct networks. However, most combined tDCS and fMRI previous studies investigated the impact of only one type of stimulation (either anodal- versus sham-tDCS or cathodal- versus sham-tDCS). Here we directly compare the impact on behavior and on neural activity of both anodal- and cathodal-tDCS, stimulation protocols that are communally accepted to have opposite effects (e.g. to induce cortical facilitation [anodal-tDCS] and inhibition [cathodal tDCS]) (Nitsche and Paulus, 2000, 2001).

To summarize, we used concurrent tDCS with fMRI to reveal the dynamic changes in functional interplay between the stimulated rLPFC and interconnected brain networks underlying norm-based social behavior. We show that cathodal-tDCS resulted in punishment-related activity blunts in regions within the executive network, consistent with the view that stimulation may have diminished strategic thinking about the consequences of the punishment threats. Importantly, anodal-tDCS increases different aspects of punishment-related neural responses (e.g. neural activity or functional connectivity) in areas within affective network, suggesting increased affective responses to punishment threats. Furthermore, our results suggests that rLPFC is not simply implementing only local neural processes which generate the appropriate decisions, but rather that rLPFC is coordinating the activity in interconnected brain networks, such as executive and affective networks, which in the end are jointly responsible for the modulation of the norm-compliant behavior. More general, our present study not only that provide more insights into the brain mechanisms underlying sanction-induced social norm compliance, but also demonstrates how concurrent combination of tDCS with fMRI can underpin the causal interplay between behavior, the stimulated site and different neural networks in the service of decision-making.

References

Amano T, Unal CT, Pare D (2010) Synaptic correlates of fear extinction in the amygdala. Nat Neurosci 13:489-494.

Antal A, Polania R, Schmidt-Samoa C, Dechent P, Paulus W (2011) Transcranial direct current stimulation over the primary motor cortex during fMRI. Neuroimage 55:590-596.

Bechara A, Damasio H, Damasio AR (2003) Role of the amygdala in decision-making. Ann N Y Acad Sci 985:356-369.

Buckholtz JW, Marois R (2012) The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. Nat Neurosci 15:655-661.

Buckholtz JW, Martin JW, Treadway MT, Jan K, Zald DH, Jones O, Marois R (2015) From Blame to Punishment: Disrupting Prefrontal Cortex Activity Reveals Norm Enforcement Mechanisms. Neuron 87:1369-1380.

Cummings JL (1995) Anatomic and behavioral aspects of frontal-subcortical circuits. Ann N Y Acad Sci 769:1-13.

Duncan J (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn Sci 14:172-179.

Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates (vol 113, pg 7900, 2016). P Natl Acad Sci USA 113:E4929-E4929.

Fehr E, Gachter S (2002) Altruistic punishment in humans. Nature 415:137-140.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218-229.

Glimcher PW, Dorris MC, Bayer HM (2005) Physiological utility theory and the neuroeconomics of choice. Games Econ Behav 52:213-256.

Hoover WB, Vertes RP (2007) Anatomical analysis of afferent projections to the medial prefrontal cortex in the rat. Brain Struct Funct 212:149-179.

Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF (2005) Neural systems responding to degrees of uncertainty in human decision-making. Science 310:1680-1683.

Levine DS (2009) Brain pathways for cognitive-emotional decision making in the human animal. Neural Netw 22:286-293.

McDonald AJ (1998) Cortical pathways to the mammalian amygdala. Prog Neurobiol 55:257-332.

Milad MR, Quirk GJ (2002) Neurons in medial prefrontal cortex signal memory for fear extinction. Nature 420:70-74.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167-202.

Moisa M, Polania R, Grueschow M, Ruff CC (2016) Brain Network Mechanisms Underlying Motor Enhancement by Transcranial Entrainment of Gamma Oscillations. J Neurosci 36:12053-12065.

Montague PR, Lohrenz T (2007) To detect and correct: norm violations and their enforcement. Neuron 56:14-18.

Morrison SE, Saez A, Lau B, Salzman CD (2011) Different time courses for learning-related changes in amygdala and orbitofrontal cortex. Neuron 71:1127-1140.

Murray EA (2007) The amygdala, reward and emotion. Trends Cogn Sci 11:489-497.

Nitsche MA, Paulus W (2000) Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. J Physiol 527 Pt 3:633-639.

Nitsche MA, Paulus W (2001) Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans. Neurology 57:1899-1901.

Nitsche MA, Cohen LG, Wassermann EM, Priori A, Lang N, Antal A, Paulus W, Hummel F, Boggio PS, Fregni F, Pascual-Leone A (2008) Transcranial direct current stimulation: State of the art 2008. Brain Stimul 1:206-223.
Ohman A (2005) The role of the amygdala in human fear: automatic detection of threat. Psychoneuroendocrinology 30:953-958.
Phelps EA, Delgado MR, Nearing KI, LeDoux JE (2004) Extinction learning in humans: role of the amygdala and vmPFC. Neuron 43:897-905.
Polania R, Nitsche MA, Ruff CC (2018) Studying and modifying brain function with non-invasive brain stimulation. Nat Neurosci 21:174-187.
Raine A, Yang Y (2006) Neural foundations to moral reasoning and antisocial behavior. Soc Cogn Affect Neurosci 1:203-213.
Ruff CC, Ugazio G, Fehr E (2013) Changing social norm compliance with noninvasive brain stimulation. Science 342:482-484.
Saddoris MP, Gallagher M, Schoenbaum G (2005) Rapid associative encoding in basolateral amygdala depends on connections with orbitofrontal cortex. Neuron 46:321-331.
Smuts B (1999) Multilevel selection, cooperation, and altruism : Reflections on unto others: The evolution and psychology of unselfish behavior. Hum Nat 10:311-327.
Spitzer M, Fischbacher U, Herrnberger B, Gron G, Fehr E (2007) The neural signature of social norm compliance. Neuron 56:185-196.
Strang S, Gross J, Schuhmann T, Riedl A, Weber B, Sack AT (2015) Be nice if you have to--the neurobiological roots of strategic fairness. Soc Cogn Affect Neurosci 10:790-796.

Curriculum Vitae
Giuseppe Ugazio

Personal Data
Date of Birth: 07.08.1986
Place of Birth: Cantú, Italy
Nationality: Italian

Education

2016 – 2018	Post-Doctoral Research Fellow, Psychology Department, Harvard University
2012 – 2018	PhD Studies, Neuroeconomics, University of Zurich
2008 – 2012	PhD, Philosophy, University of Zurich (*Magna cum laude*)
2007 – 2008	MSc, Philosophy of the Social Sciences, London School of Economics
2004 – 2007	BA, Philosophy, Università Vita-Salute San Raffaele, Milan

Professional Experience

2018 – Present	**Harvard University - Lecturer**, *Evolution of Human Cooperation*, Human Evolutionary Biology Department.
2016 – Present	**Harvard University - Post-Doctoral Research Fellow**, Prof. Fiery Cushman's Moral Psychology Research Lab.
2017	**Harvard University - Teaching Fellow**, *Hormones and Behavior*, Human Evolutionary Biology Department.
2017	**Harvard University - Teaching Fellow**, *Evolving Morality*, Psychology Department.
2016	**Harvard University - Teaching Fellow**, *Human Nature*, Human Evolutionary Biology Department.
2014 – 2016	**University of Zurich - Post-Doctoral Research Fellow**, Professors Christian Ruff and Ernst Fehr's Social and Neural Systems Research Lab.
2013 – 2014	**University of Vienna - SNSF Early-Career Post-Doctoral Research Fellow**, Prof. Claus Lamm's Social Cognitive and Affective Neuroscience Unit.